DOCUMENT RESUME

ED 338 627 TM 012 520

TITLE	Military Testing Association (Proceedings of the Annual Conference (27th, San Diego, California, October 21-25, 1985). Volume I and Volume II.
INSTITUTION	Navy Personnel Research and Development Center, San Diego, Calif.
PUB DATE	Oct 85
NOTE	1,077p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF08/PC44 Plus Postage.
DESCRIPTORS	*Military Personnel; *Research Methodology; *Testing

ABSTRACT

The 169 paper and symposium presentations given during 57 sessions of the conference are provided in these two volumes. The first volume contains the keynote speech, which addressed military entrance processing command and its acquired immune deficiency snydrome testing program. Symposia topics in this volume include: (1) computerized diagnostic testing in the Navy's basic electricity and electronics school; (2) computerized adaptive testing research at the Navy Personnel Research and Development Center; (3) systems for controlling performance to improve productivity; (4) Air Force training decisions systems research; and (5) issues in job performance measurement. Paper session topics include: (1) training evaluation; (2) computerized adaptive testing; (3) leadership dimensions; (4) research implementation; (5) statistical methods in test construction; (6) uses of microcomputers in testing; (7) officer career development; (8) computer-assisted assessment of human performance; (9) manpower research; (10) computers and videos in instruction and training assessment; (11) occupational analysis; (12) organizational effectiveness; (13) computer-based testing; (14) occupational data analysis systems; (15) behavioral aspects of performance; (16) computer-based instruction; and (17) job evaluation. In the second volume, symposia topics include: (1) remedial education in the Navy; (2) elements of a Military Occupational Exploration System; (3) determining ability requirements; (4) exceptionally low- and high-aptitude recruits; (5) predicting a broad variety of criteria by elaborating the predictor space; (6) quanitatiave and qualitative approaches to the training and selection of Army managers; (7) Army Research Institute research and development program on the National Training Center; and (8) two systems for computer analysis of student feedback. Paper session topics include: (1) basic skills training; (2) survey methodology and technology; (3) biodata use and validation; (4) training device evaluation; (5) attitude surveys; (6) selection methodology; (7) team performance measurement; (8) analysis of the Armed Services Vocational Aptitude Battery (ASVAB); (9) personality assessment; (10) team training; (11) integrated selection systems; (12) test item design; (13) entry standards based on the ASVAB; (14) military attrition research; (15) officer training and testing; (16) job performance measurement; (17) instructional issues; and (18) assessment of English-as-a-Second-Language personnel. MTA bi-laws, minutes of the Steering Committee Meeting, a description of the Harry H. Greer Award, lists of MTA conference staff members and registrants, and an author index are appended. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
K.E. WILCKEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PROCEEDINGS

27th Annual Conference of the

MILITARY TESTING ASSOCIATION

Coordinated by the NAVY PERSONNEL RESEARCH & DEVELOPMENT CENTER

SAN DIEGO, CALIFORNIA
21 - 25 OCTOBER 1985

Volume I

BEST COPY AVAILABLE

ED338627
TM 017520

PROCEEDINGS

27TH ANNUAL CONFERENCE

of the

MILITARY TESTING ASSOCIATION

Coordinated by

the

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

SAN DIEGO, CALIFORNIA

21-25 OCTOBER 1985

VOLUME I

FOREWORD

The papers presented at the Twenty-Seventh Annual Conference of the Military Testing Association came from the business, educational, and military communities, both foreign and domestic. The papers reflect the opinions of their authors only and are not to be construed as the official policy of any institution, government, or branch of the armed services.

CONTENTS

VOLUME I

SESSION	TITLE	AUTHOR	PAGE
	Foreward		ii
OPENING:	**Call To Order**	Wiskoff, M.F.	xxvi
	Welcome	Eldredge, CAPT H.S.	xxvi
	Keynote: The U.S. Military Entrance Processing Command and its Testing Program for the AIDS Virus	Archer, BG C.J.	xxvi
SYMPOSIUM:	**Computerized Diagnostic Testing in the Navy's Basic Electricity and Electronics School** Chairperson: Davis, C.E. Office of Naval Research		
	Rule Space: A Model for Identifying Erroneous Rules of Operation	Tatsuoka, M.	1
	Implementation of a Computer System to Support Diagnostic Testing	Vale, C.D.	7
	Application of Rule Space in a Navy Testing Environment	Eddins, J.M.	13
SYMPOSIUM:	**Computerized Adaptive Testing Research the Navy Personnel Research and Development Center** Chairperson: Sands, W.A. Navy Personnel Research and Development Center		
	An Overview of the Accelerated CAT-ASVAB Program	Sands, W.A.	19
	Evaluation of Three Local CAT-ASVAB Network Designs	Tiggle, R.B. Rafacz, B.A.	23
	A Validity Study of the Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery	Moreno, K.E. Segall, D.O. Kieckhaefer, W.F.	29

SESSION	TITLE	AUTHOR	PAGE
	Assessment of the Unidimensionality of CAT-ASVAB Subtests	Schratz, M.K.	34
	Sex Differences in IRT True-Score Equating	Segall, D.O. Kieckhaefer, W.F. Moreno, K.E.	38
	Reducing the Predictability of Adaptive Item Sequences	Wetzel, C.D. McBride, J.R.	43
	Speeded Tests - Can Computers Improve Measurement?	Wolfe, J.H.	49
	Medium of Administration Effects on Attitudes toward ASVAB Testing	Kieckhaefer, W.F. Segall, D.O. Moreno, K.E.	55
SYMPOSIUM:	**Systems for Controlling Performance to Improve Productivity** Chairperson: Dockstader, S.L. Navy Personnel Research and Development Center		
	Statistical Process Controls as an Enhancement to Job Design	Dockstader, S.L.	61
	A Group Wage Incentive System: Design and Implementation Issues	Mohr, D.A.	67
	Goal Setting with Navy Production Workers: Problems and Potential	Crawford, K.S.	73
	The Determinants of Goal Choice, Work Motivation and Task Performance	Riedel, J.A. Nebeker, D.M.	78
	The Effects of Reward Magnitude and Difficulty of Performance Standards upon Individual Productivity	Nebeker, D.M.	82

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Training Evaluation I** Chairperson: Van Matre, N. Navy Personnel Research and Development Center		
	Results of an External Evaluation System Development in the Canadian Forces	Kerr, CDR R.H. Tyerman, D.	88
	Occupational Learning Difficulty: A Construct Validation against Training Criteria	Weeks, J.L. Mumford, M.D. Harding, F.D.	90
PAPER SESSION:	**Developments in Computerized Adaptive Testing** Chairperson: Schratz, M.K. Navy Personnel Research and Development Center		
	News about CAT in the German Federal Armed Forces(GFAF)	Wildgrube, W.	96
	Efforts towards the Improvement of the Computerized Adaptive Screening Test (CAST)	Knapp, D.J. Pliske, R.M. Johnson, R.M.	102
	Development of a Portable Computerized Performance Test System	Kennedy, R.S. Dunlap, W.P.. Wilkes, R.L. Lane, N.E.	107
PAPER SESSION:	**Dimensions of Leadership I** Chairperson: Atwater, D.C. Navy Personnel Research and Development Center		
	Leadership Preparedness of Newly Commissioned Naval Officers	Atwater, L. Crawford, A.	113
	Leadership Development of USAF Aircraft Maintenance Officers	Morabito, CPT M.A. Dilla, CPT B.L.	118

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Research Implementation** Chairperson: Van Matre, N. Navy Personnel Research and Development Center		
	Implementing Technology in a Military Environment	Maloy, W.L. Perry, N.N.	124
	Research Needs Assessment and Technology Transfer in USAREUR	Girdler, K. McLain, LTC F.	130
PAPER SESSION:	**Statistical Methods in Test Construction** Chairperson: Sympson, J.B. Navy Personnel Research and Development Center		
	The Reduction of Standard Errors of Equipercentile Test Equating through Negative Hypergeometric Presmoothing	Fairbank, B.A.	136
	Weighted Likelihood Estimation of Ability in Item Response Theory with Tests of Finite Length	Warm, T.A.	142
PAPER SESSION:	**Dimensions of Leadership II** Chairperson: Atwater, D.C. Navy Personnel Research and Development Center		
	Leader Behavior and the Performance of First Term Soldiers	White, L.A. Gast, I.F. Rumsey, M.G.	147
	Measuring Military Heriosm	Anderson, J.W.	153

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Training Evaluation II** Chairperson: Van Matre, N. Navy Personnel Research and Development Center		
	Analysis of Functional Training Programs at the Naval Supply Centers	Crowley, N.R.	159
	Systematic Curriculum Review	Gottesman, A.M.	165
PAPER SESSION:	**Uses of Microcomputers in Testing** Chairperson: Tiggle, R.B. Navy Personnel Research and Development Center		
	Computer Aversion as a Source of Bias in Computerized Testing	Wood, J.A. Pitz, G.F.	171
	Orientation, Surrogate Travel, and Gender Differences in Videogame Strategy	Tkacz, S.	177
PAPER SESSION:	**Officer Career Development I** Chairperson: Bruni, J.R. Navy Personnel Research and Development Center		
	The Navy's General Unrestricted Line Community: Career Management and Career Development Problems	Wilcove, G.L.	183
	Attitudes, Preferences and Career Intentions of ROTC and Non-ROTC Students	Kimmel, M.J.	188

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Training Evaluation III** Chairperson: Thode, W.F. Navy Personnel Research and Development Center		
	Can Maintenance Data Be Used to Define Training Requirements?	Wagner, M. Costellic, MAJ M.	194
	Training Emphasis: The Best Task Factor Available	Hardy, LT D.L. Gorman, LTC C.D. Driskill, W.E.	198
	Training Effectiveness Analysis of DRAGON Simulators	Butler, W.G.	204
PAPER SESSION:	**Officer Career Development II** Chairperson: Chatfield, R.E. Navy Personnel Research and Development Center		
	Career Involvement of Recent West Point Graduates	Adams, J. Richards, J.D.	210
	Mentoring in the United States Air Force: The Mentor's Perspective	Lewandowski, CPT F. Dilla, CPT B.L.	216
	Preparing to Evaluate Career Path Changes: Pre-change Database Development	Chatfield, R.E. Morrison, R.F.	222
PAPER SESSION:	**Training: General Topics** Chairperson: Thode, W.F. Navy Personnel Research and Development Center		
	USAF Officer Professional Military Education	Bell, 1LT J.M. Tartell, J.S.	228

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Computer Assisted Assessment of Human Performance Characteristics** Chairperson: Wolfe, J.H. Navy Personnel Research and Development Center		
	Visual, Auditory, and Reaction Time Measurement of Speed of Information Processing and Intelligence	Saccuzzo, D.P. Larson, G.E. Rimland, B.	233
	Representation and Modification of Human Performance Factors in Target Selection	Goehring, D.J.	239
PAPER SESSION:	**Manpower Research** Chairperson: Chatfield, R.E. Navy Personnel Research and Development Center		
	Quality of Canadian Forces Other-rank Recruits: Trends 1975-1983	Wenek, MAJ K.W.J.	245
	Role, Importance, and Availability of GFAF-Reservists	Ebenrett, H.	251
	Retirement Readiness as a Function of Transition Assistance and Trade	Wilson, MAJ F.P.	256
SYMPOSIUM:	**Air Force Training Decisions Systems Research** Chairperson: Garcia, S.K. Air Force Human Resources Laboratory		
	Development of an Air Force Training Decisions System	Garcia, S.K.	262
	Defining Task Training Modules: Coperformance Clustering	Perrin, B.M. Vaughan, D.S. Yadrick, R.M. Mitchell, J.L.	265

SESSION	TITLE	AUTHOR	PAGE
	Evaluating Task Training Modules: SME Clustering and Comparisons	Yadrick, R.M. Vaughan, D.S. Perrin, B.M. Mitchell, J.L.	271
	Nonhierarchial Clustering of Air Force Jobs and Tasks	Mitchell, J.L. Phalen, W.J.	276
SYMPOSIUM:	**Issues in the Development of CAT-ASVAB** Chairperson: Wainer, H. Education Testing Service		
	On the Study of Differential Item Performance without IRT	Holland, P.W.	282
	Reviewing an Item Pool for its Sensitivity to the Concerns of Women and Minorities: Process and Outcomes	Kershaw, S.W. Wainer, H.	288
	Measurement Precision and "Reliability": Some Considerations of Metrics and Stopping Rules in CAT	Thissen, D.	294
SYMPOSIUM:	**Issues in Job Performance Measurement** Chairperson: Laabs, G.J. Navy Personnel Research and Development Center		
	Multi-dimensional Performance Measurement	Hanser, L.M. Arabian, J.M. Wise, L.	300
	Walk-through Performance Test Development: Lessons Learned	Taylor, C.J. Blackhurst, J.L. Ballentine, R.D.	306
	On the Content and Measurement Validity of Hands-on Job Performance Tests	Maier, M.H.	311

SESSION	TITLE	AUTHOR	PAGE
	Developing and Evaluating a Hands-on Performance Test	Kroeker, L.P. Bearden, R.M. Laabs, G.J.	317
	Simulation of Hands-on Testing for Navy Machinist's Mates	Vineberg, R. Joyner, J.N.	323
	Inter-Service Transfer of Job Performance Measurement Technology	Blackhurst, CPT J.L. Baker, H.G.	327
PAPER SESSION:	**Computers and Video in Instruction and Training Assessment** Chairperson: McDonald, B.A. Navy Personnel Research and Development Center		
	Simulation of Instructor and Group Process Roles with Microcomputer Technology	McCombs, B.L.	331
	Assessing M1 Tank Commanders with a Computerized Hand-held Tutor	Bridgeman, B. Post, T.	337
	Microform in Training	Jones, LCDR K. Mizen, LCDR A.E.	341
PAPER SESSION:	**Occupational Analysis** Chairperson: Baker, H.G. Navy Personnel Research and Development Center		
	Selling Job Analysis	Stephenson, S.D.	346
	Linking Work, Training, and Promotion in the Coast Guard	Jones, K.N. Burt, J.A..	352
	Determining the Best Set of Training Factors for Use with Army Occupational Surveys	Goldman, L.A.	358

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Organizational Effectiveness** Chairperson: Thomas, P.J. Navy Personnel Research and Development Center		
	Examination of Environmental Determinants of Army Performance Criteria	Olson, D.M. Borman, W.C.	367
	Initial Standardization of an Air Force Organizational Assessment Survey Instrument	Short, LTC L.O. Lowe, CPT J.K. Hightower, CPT J.M.	375
	Productivity Measurement: Issues and Challenges	van Rijn, P.	381
	A Morale and Mission Relating Strategy	Waldkoetter, R.O.	387
PAPER SESSION:	**Computer-Based Testing** Chairperson: McDonald, B.A. Navy Personnel Research and Development Center		
	COGENT - A Microcomputer Based Criterion Referenced Testing System	Kerkman, D.	393
	Microcomputer Based Field Testing for Human Performance Assessment	Merkle, P.J. Kennedy, R.S. Smith, M.G. Johnson, J.H.	398
	Applications of a Portable Training/ Testing Computer	Harman, J.	404
PAPER SESSION:	**Occupational Data Analysis Systems** Chairperson: Folchi, J.S. Navy Personnel Research and Development Center		
	Design of an Occupational Data Analysis System	Wheeler, MAJ C.P.	408

SESSION	TITLE	AUTHOR	PAGE
	Implementation of Air Force ASCII CODAP: New Versus Old System	Phalen, W.J. Staley, M.R. Weissmuller, J.J.	414
	Occupational Research Data Bank(ORDB)	Ellingsworth, M.E. Olivier, L.F. Pfeiffer, G.J.	420
PAPER SESSION:	**Behavioral Aspects of Performance** Chairperson: Thomas, P.J. Navy Personnel Research and Development Center		
	Effects of Absenteeism on the Performance of Air National Guard Personnel	Steel, R.P. Shane, G.S.	426
	In Search of Honesty	Shub, A.N.	432
	A Model of Psychological Stress and Control	Kastner, M. Neef, W.	438
PAPER SESSION:	**Computer-Based Instruction** Chairperson: McDonald, B.A. Navy Personnel Research and Development Center		
	Methodological Considerations in CBI Research	Shlechter, T.M. Boldovici, J.A.	444
	The Effectiveness of Computer Graphics in Visual Recognition Training	McDonald, B. Whitehill, B.	450
PAPER SESSION:	**Job Evaluation** Chairperson: Robertson, D.W. Navy Personnel Research and Development Center		
	Recent Developments in Job Evaluation Research	Stephenson, S.D.	456

CONTENTS

VOLUME II

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Basic Skills Training** Chairperson: Konoske, P.J. Navy Personnel Research and Development Center		
	Pre-Course Mathematics Skills and Success on Advanced Naval Technical Courses	Amyot, CAPT K.A.	462
	Causes of Performance in Air Force Initial Skills Training	Harding, F.D. Mumford, M. Weeks, J.L.	467
PAPER SESSION:	**Methodological and Technological Developments in Survey Research** Chairperson: Morrison, R.F. Navy Personnel Research and Development Center		
	An Examination of the Significance of Non-Participation in Field Studies	Potter, E.H.	473
	Advancing Survey Design through Technology	Doherty, L.M.	480
	Implementation of a Civilian Automated Survey System	Ripkin, F.L. Cecil, S.	486
PAPER SESSION:	**Utilization and Validation of Biodata** Chairperson: Mattson, J.D. Navy Personnel Research and Development Center		
	The Utility of Educational and Biographical Information for Predicting Military Attrition	Riegelhaupt, B.J. Bonczar, T.P.	491

SESSION	TITLE	AUTHOR	PAGE
	Three Variables that May Influence the Validity of Biodata	Walker, C.B.	497
	Use of Personal History Information to Predict Naval Academy Disenrollment	Mattson, J.D. Abrahams, N.M. Hetter, R.D.	503
	Intercorrelations of Biographical Information, Aptitude Test Scores, and Job Performance Ratings for 108 Occupations	Trattner, M.H.	509
PAPER SESSION:	**Training Device Evaluation** Chairperson: Konoske, P.J. Navy Personnel Research and Development Center		
	Analytic Prediction of Training Device Effectiveness	Yates, L.G. Macpherson, D.	515
	Research on Decision Aids for Training Design and Evaluation	Mirabella, A.	520
PAPER SESSION:	**Attitude Surveys** Chairperson: Wilcove, G.L. Navy Personnel Research and Development Center		
	The Army Experience Survey: Methodological Highlights	Celeste, J.F.	526
	USAF Spouse Survey: The Final Chapter	Dansby, MAJ M.R. Ibsen, CPT K.A.	531
	Predictors of Propensity for Continuing Education among Army Chaplains	Pierce, J.E. Goldman, L.A.	537

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Selection Methodogy** Chairperson: Moreno K.E. Navy Personnel Research and Development Center		
	Composite Ratings as a Performance Criterion	Warren, J. Newton, P. Bondaruk, J.	543
	Maximizing Criterion Variance in Validation Research: Is This Always Best?	Angus, CAPT R.J. Ellis, MAJ R.T.	547
SYMPOSIUM:	**Remedial Education in the Navy** Chairperson: Chang, F.R. Navy Personnel Research and Development Center		
	Literacy, Readability and Knowledge	Chang, F.R.	553
	The XFSP: Reading and Mathematics Project	Sticht, T. Armijo, L. Koffman, N. Roberson, K.	559
SYMPOSIUM:	**Elements of a Military Occupational Exploration System** Chairperson: Pass, J.J. Navy Personnel Research and Development Center		
	A Conceptual Framework for Occupational Exploration	Pass, J.J.	565
	Development of the Career Maturity Assessment	Diamond, E.E.	570
	Navy R&D in Support of the Army JOIN System: Leveraging the Government Research Dollar	Baker, H.G. Rafacz, B.A. Sands, W.A.	576

SESSION	TITLE	AUTHOR	PAGE
	Vocational Interests as Predictors of Army Performance	Wing, H. Barge, B.N. Hough, L.M.	582
	Army Career Vocational Guidance as a Recruiting Tool	Hertzbach, A. Knapp, D.J. Johnson, R.M.	587
SYMPOSIUM:	**Determining Ability Requirements** Chairperson: Abrabian, J.M. U.S. Army Research Institute		
	Developing New Attribute Requirements Scales for Military Jobs	Smith, E.P.	593
	Methodological Problems in Identifying Ability Requirements Related to Soldier Performance	Miller, C.R.	599
	Comparison of Weapon Systems Using Ability Requirements Scales	Arabian, J.M.	603
	Computerized Approaches for Estimating Ability Requirements	Rossmeissl, P.G.	609
PAPER SESSION:	**Team Performance Measurement** Chairperson: Flaningam, M.R. Navy Personnel Research and Development Center		
	Chaparral Crew Performance in the Realistic Air Defense Engagement System	Sarli, G.G. Johnson, D.M. Lockhart, J.M.	615
	Command and Control Teams: Techniques for Assessing Team Performance	Cooper, M. Shiflett, S. Korotkin, A.L.	621

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Analysis of the ASVAB** Chairperson: Larson, G.E. Navy Personnel Research and Development Center		
	Item Factor Analysis of ASVAB 14	McCormick, C.	627
	Alternate Forms Reliability of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10	Stern, B.M. White, L.A. Wing, H. Sachs, S.A.	632
PAPER SESSION:	**Personality Assessment** Chairperson: Bjerke, CDR D.G. Navy Personnel Research and Development Center		
	Psychometric Properties of the Safety Locus of Control Scale	Jones, J.W. Wuebker, L.J.	637
	Health Locus of Control Beliefs among Infantrymen	Grieger, L.W.	644
	Classifying Military Offenders: Application of the Megargee MMPI Typology	Paris, M.L. Brown, G.E.	650
PAPER SESSION:	**Team Training** Chairperson: Flaningam, M.R. Navy Personnel Research and Development Center		
	Automated Scheduling of Army Unit Training	Goehring, D.J. Hart, R.J.	655
	Supporting the Changing Role of Army Collective Training Developers	Meliza, L.L.	661
	Application of Model Aircrew Training System (MATS) to B-52 Combat Crew Training	Bills, MAJ C.G. Nullmeyer, R.T.	666

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Analysis and Interpretation of the ASVAB** Chairperson: Moreno, K.E. Navy Personnel Research and Development Center		
	Sensitivities of Speeded Subtests	Wegner, T.G. Ree, M.J.	672
	Initial Operational Test and Evaluation of Armed Service Vocational Aptitude Battery (ASVAB) Forms 11, 12, and 13: Data Quality Analysis	Welsh, J.R. Wegner, T.G.	676
PAPER SESSION:	**Integrated Selection Systems** Chairperson: Bruni, J.R. Navy Personnel Research and Development Center		
	Development of an Integrated Pilot Selection System	Kantor, J.E.	680
	Selection of Skilled Maintenance Employees in the U.S. Postal Service	Mueller, H.A.	686
	The Effects of the Flight Screening Program on Attrition in Undergraduate Pilot Training	Quebe, J.C.	695
PAPER SESSION:	**Issues in Test Item Design** Chairperson: Flaningam, M.R. Navy Personnel Research and Development Center		
	What Performance Does a Performance Test Test?	Ansbro, T.M.	701
	How Qualitatively Informative are Test Items?: A Dense Item Analysis	Bart, W.M.	707
	Test-item Readability: How the Variables Work	Duncan, CPT R.E.	713

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Entry Standards Based on the ASVAB** Chairperson: Vicino, F.L. Navy Personnel Research and Development Center		
	A Comparison of Service Job Standards for Four Military Specialties	Waters, B.K.	719
	Exploring a Statistically Viable Assignment Basis Using ASVAB	Schwartz, M.M.	725
	The Validity of ASVAB for Predicting Training and SQT Performance	Rossemeissl, P.G. McLaughlin, D.H. Wise, L.L. Brandt, D.A.	730
PAPER SESSION:	**Military Attrition Research** Chairperson: Cory, C.H. Navy Personnel Research and Development Center		
	Gender, Ethnic Group, Aptitude and Personality Determinants of U.S. Coast Guard Attrition	Frey, R.L.	736
	Study of Wastage from the Territorial Army (UK Army Reserve)	Blyth, D.M.	741
	Diminishing Returns from an Assessment Centre	Hardy, G.R.	747

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Officer Training and Testing** Chairperson: Greebler, C.S. Navy Personnel Research and Development Center		
	Increasing Basic Skill Levels of Moderate Risk Officer Candidates Prior to Commissioning in Historically Black Colleges	Sprenger, W.D.	753
	Direct Mesurement of ROTC Cadets' Writing Skills	Hanlon, J.P.	759
	Procedures for Refining Written Measurements at USAF OTS	Sako, S. Slaughter, LTC W.J.	763
PAPER SESSION:	**Job Performance Measurement I** Chairperson: Kroeker, L.P. Navy Personnel Research and Development Center		
	Utility Estimation in Five Enlisted Occupations	Eaton, N.K. Wing, H. Lau, A.	769
	Relation of Mental and Education Levels to Navy Enlisted Performance	Cory, C.H.	775
	Selecting Critical Tasks for a Radioman Hands-on Performance Test	Baker, H.G. Laabs, G.J.	780
	Cognitive Predictors of M1 Tank Gunner Performance	Black, B.A.	786

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Issues in Instruction I** Chairperson: Windisch, D.J. Navy Personnel Research and Development Center		
	A Retrospective Analysis of Instructional Technology Innovation Using General System Theory	Ekstrom, A.G.	792
	A Systems Approach to Evaluating High Technology Training	Atwood, N.K.	798
	An Analysis of Attitudes toward Instruction among Vocational Education Instructors	Usova, G.M.	804
PAPER SESSION:	**Assessment of English-as-a-Second Language Personnel** Chairperson: Hetter, R.D. Navy Personnel Research and Development Center		
	Relationship of an Experimental Hispanic Enlistment Screening Test to AFQT	Mathews, J.J. French, C.M.	809
	Attrition and Performance Ratings of ESL Soldiers in BT	Rosenbaum, H.	815
PAPER SESSION:	**Job Performance Measurement II** Chairperson: Kroeker, L.P. Navy Personnel Research and Development Center		
	Standard Setting Methods for Skills Qualification Tests (SQTs)	Pettie, A.L.	821
	Assessing Tank Commander and Gunner Proficiency on U-COFT	Graham, S.E. Boldovici, J.A.	826
	Criterion Referenced Testing for the U.S. Navy's Nuclear Submarine Fleet	Cantor, J.A. Walker, L.	832

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Issues in Instruction II** Chairperson: Windisch, D.J. Navy Personnel Research and Development Center		
	A Study of Suggestopedia at the Defense Language Institute Foreign Language Center (DLIFLC)	Bush, B.J.	838
	Heat, Chemical Protective Clothing and Sustained Cognitive Performance	Fine, B.J. Kobrick, J.L.	843
SYMPOSIUM:	**Exceptional Recruits: A Look at High and Low Aptitude Personnel** Chairperson: Lancaster, A.R. Office of the Assistant Secretary of Defense (FM&P)		
	Overview	Lancaster, A.R.	849
	Using Military High- and Low- Aptitude Recruits Wisely	Sellman, W.S.	852
	When Low-Aptitude Recruits Succeed	Means, B. Nigam, A. Heisey, J.G.	855
	Enlistment and Utilization of High-Aptitude Recruits	Laurence, J.H. Schneider, E.F.	861
SYMPOSIUM:	**Predicting a Broad Variety of Criteria: Elaborating the Predictor Space** Chairperson: Wing, H. U.S. Army Research Institute		
	Mapping Predictors to Criterion Space: Overview	Peterson, N.G.	867
	Using Microcomputers for Assessment: Practical Problems and Solutions	Rosse, R.L. Peterson, N.G.	873

SESSION	TITLE	AUTHOR	PAGE
	Computerized Assessment of Perpetual and Psychomotor Abilities	McHenry, J.J. Toquam, J.L.	879
	Adding to the ASVAB: Cognitive Paper-and-Pencil Measures	Toquam, J.L. Dunnette, M.D. Corpe, V.A. Houston, J.	885
	Measuring Personnel Attributes: Temperament, Biodata and Interest	Hough, L.M. McGue, M.K. Kamp, J.D. Houston, J.S. Barge, B.N.	891
SYMPOSIUM:	**The Training and Selection of Army Managers: Quantitative/Qualitative Approaches** Chairperson: Fisher, G.P. Human Resources Research Organization		
	Overview	Fisher, G.P. Lilienthal, R.	897
	Selecting and Training Logistics Professionals and Managers: Qualitative/ Quantitative Approaches	Fisher, G. Lilienthal, R. Hough, L.	900
	Managerial Competencies Assessment For Army Civilians	King, G. C.	906
SYMPOSIUM:	**Army Research Institute R&D Program on the National Training Center** Chairperson: Banks, J.H. U.S. Army Research Institute		
	Types and Quality of National Training Center Data	Whitmarsh, P.J.	912
	An Overview of ARI's Research Program on the National Training Center	Banks, J.H.	916

SESSION	TITLE	AUTHOR	PAGE
	Leader Performance Criteria at the National Training Center (NTC)	Pence, E.C.	922
	Battalion Performance on the Live Fire Range at the National Training Center (NTC)	Forsythe, T.K. Doherty, W.J.	928
	Analysis of NTC Force-on-Force Performance	Nichols, J.J. Doherty, W.J.	931
	Use of Instrumentation to Improve Combat Readiness	Shackelford, W.L.	937
SYMPOSIUM:	**Two Systems for Computer Analysis of Student Feedback** Chairperson: Ekwall, R.W. U.S. Army Command and General Staff College		
	Collecting and Utilizing Feedback: Combined Arms and Services Staff School	Ekwall, R.W.	943
	By-laws of the Military Testing Association		949
	Minutes of the Steering Committee		953
	Harry H. Greer Award		956
	MTA 27th Annual Conference Staff		957
	Military Testing Association Conference Registrants		958
	Author Index		970

OPENING SESSION OF THE 27TH ANNUAL MILITARY TESTING ASSOCIATION CONFERENCE

21 October 1985

The 27th Annual Conference of the Military Testing Association was hosted by the Navy Personnel Research and Development Center (NPRDC). The Conference was held at the Bahia Hotel in San Diego, California, 21 through 25 October 1985. A total of 169 paper and symposium presentations were given during 57 sessions structured into three concurrent tracks. Conference attendance was 326.

CALL TO ORDER: Dr. Martin F. Wiskoff, Head, Manpower and Personnel Laboratory, NPRDC, and MTA Chairman called the Conference to order at 1300, 21 October. Dr. Wiskoff then introduced the MTA President, CAPT Howard S. Eldredge, Commanding Officer, NPRDC.

WELCOME: CAPT Eldredge officially welcomed the attendees to the Conference and San Diego. He discussed the importance of personnel systems to missions of the armed forces and emphasized how they represent the key to military superiority. CAPT Eldredge then expressed the hope that the formal presentations and the informal interchanges the attendees would experience during the week would go far to further personnel research and development.

KEYNOTE: Prior to the introduction of the keynote speaker, Brigadier General Caleb J. Archer, Commander, U.S. Military Entrance Processing Command, Dr. Wiskoff gave a brief summary of the General's career. Dr. Wiskoff noted that General Archer has held a wide variety of important command and staff positions. In 1966 he served in Vietnam commanding the 212th Military Police Company and later was Chief of Physical Security for the U.S. Army Vietnam. General Archer also served on the faculty of the Army Command and General Staff College and has served as Provost Marshal of the 3rd Infantry Division in Wurzburg, Germany and Commander of the 793rd Military Police Battalion, Nuremburg, Germany. General Archer has also served in Washington, DC as the Army Deputy Chief of Staff for Personnel. He has been the Provost Marshal of the Army Field Artillery Center at Fort Sill, Oklahoma. His most recent assignments have included Commander, Western Region Recruiting Command at the Presidio of San Francisco, Commandant of the Army Military Police School and Deputy Commanding General of the Army Military Police and Chemical Training Centers, Fort McClellan, Alabama

Dr. Wiskoff then asked the attendees to give a warm welcome to General Archer. (General Archer's address follows in its entirety.)

THE U. S. MILITARY ENTRANCE PROCESSING COMMAND AND ITS TESTING PROGRAM FOR THE AIDS VIRUS

Brigadier General Caleb J. Archer
U. S. Military Entrance Processing
Command

CAPT Eldredge, Dr. Lancaster, Dr. Wiskoff, ladies and gentlemen.

It is a distinct privilege and honor for me to be invited to speak to this group of researchers and professionals. Although I've only been the Commander of the U. S. Military Entrance Processing Command (USMEPCOM) for three months, I've seen already what an important part research plays in our command and, in fact, the whole Armed Forces. Your efforts in developing the various forms of testing are very important to us. As past Commandant of the Military Police School I can also assure you of the importance of other areas of personnel research and development such as your efforts involving women in the Armed Services.

At this time we in USMEPCOM are faced with the complex issue of the Acquired Immune Deficiency Syndrome (AIDS). It is an issue that will require immense effort to see the problem to a successful solution. We'll need your help.

What I would like to do today is first to tell you a little bit about the Command and then to talk about the mission we have received in USMEPCOM to test all new applicants to the Armed Forces for the HTLV-III antibody related to the AIDS virus.

USMEPCOM reports directly to the Department of Defense through Dr. Steve Sellman and Dr. Lancaster's office, and up to LTG Chavarrie. We have a very direct route through Headquarters DCSPER because they are the executive agent for processing for the Armed Forces. Our biggest customers are the four recruiting commands - Army, Navy, Air Force, and Marine Corps - and the fourteen training centers - eight Army, three Navy, 1 Air Force, and 2 Marine Corps. We qualify applicants for this complex system.

We have 68 Military Entrance Processing Stations (MEPS) throughout the U. S. and overseas and two substations - one in Guam and one in Alaska. This structure is broken down into three sectors - the Western Sector under Marine Corps COL Bill Stroup, headquartered in San Francisco; the Central Sector under Army COL James Tyler, headquartered in Chicago; and the Eastern Sector under Air Force COL Linda Sendt at Fort Meade, Maryland. We are a joint staff agency with about 49%, Army, 21% Navy, 19% Air Force, and 11% Marine Corps. The service affiliations of our commanders reflect these service percentages. However, we have 24 female commanders out of a total of 68 because all of our Navy commanders are female. This is because the Navy is seeking command billets for this level of female officer and, fortunately for us, we are able to take advantage of the situation.

Our screening process is based on the Armed Forces Vocational Aptitude Battery (ASVAB). We administer the ASVAB to about one million applicants at our testing sites each year and we administer an additional million in high schools. So we give the test about two million times each year.

We give about 540,000 physical exams at our stations a year but next year that will rise to 700,000 because we are taking on the responsibility of examining for the Army National Guard starting 1 December.

In addition to testing, USMEPCOM interviews applicants to detect law violations and other disqualifying past conduct.

Out of the total, the selection process pares the number down to about 380,000 that actually join the Armed Forces through the Military Entrance Processing Stations (MEPS) per year. All but about 5% of the total enter the Delayed Entry Program. This program allows the participants to delay entrance for up to a year. Currently the program has about 130,000 in it but this number varies up to about 150,000. So we are always working with the future in mind in our enlistment process.

For our production tests, the ones we give to the million applicants, we use the ASVAB 11, 12, and 13 forms which many of you have worked on. For the school test, administered in the high schools and junior colleges, we use ASVAB 14. With the help of Dr. Lancaster and others, we have just got testing specialists in all 81 of our stations. These specialists help market ASVAB to the schools. We market it as the excellent counseling tool it is but, of course, it also has great value to the recruiters. It allows the recruiters to direct their attention to the high quality youngsters the military needs to enlist.

We have some difficulty getting into some of the high schools to test. We give the ASVAB in about 79% of the schools but some of them test only three to five examinees so the totals don't add up very fast. We only test about 13% of the available high school juniors and seniors in the country. Endorsements from state agencies, city officials, and school administrators help the program, but the strong feeling that any outside activity detracts from the education process plays against our initiatives.

We also administer special purpose tests at the time of enlistment for the services - tests which many of you helped develop - and my staff and I participate in the planning of Department of Defense research.

I now want to give you an example of the kind of research most of our commanders grasp at -- specifically that which gives the commander a product in the near term. When we marketed the ASVAB to the schools in July 1984, we thought the test would take only three hours to administer but with the inclusion of the instructions the time was increased to about three hours and twenty minutes. Now this does not sound like much of a problem but all

the testing periods had been blocked into three hours. The school bells rang and students got up and wandered off. We had many incompleted tests and test security was compromised. Some of the schools quit the program and the recruiters were becoming upset.

A research program was quickly implemented by the Command with Navy trainees to see if shortened instructions would still be effective. It was determined that both an experimental and control group did equally as well and that shortening the instructions would have no impact on test validity. So we were able to shorten the test back to three hours and continue with the high school testing program.

Another example of research with great value to USMEPCOM is the Computerized Adaptive Testing - Armed Services Vocational Aptitude Battery (CAT-ASVAB) program which will be discussed thoroughly this week. CAT is an example of basic research which has matured to the practical level; we will be putting it to use in the near future.

I won't cover the CAT-ASVAB milestones since they will be discussed in depth during the week's sessions. However, the one I am interested in is when we declare victory and actually start CAT-ASVAB testing in the MEPS. In the meantime it is going to be a burden to get CAT-ASVAB evaluated, especially for the recruiters who must bring in applicants and tell them they will have to take 6 hours of tests. We are going to pay the price to stay on schedule. Already we are getting resistance from the recruiter force but CAT-ASVAB will be operationalized.

What I would like to do now is to switch to a discussion of our effort to test for the AIDS virus. Most of us believed up until a few months ago that this was a disease that affected homosexual men, intravenous drug users, and, tragically, hemophiliac patients but did not have much impact on the population in general. But I can tell you now that an estimated 1.2 million Americans are currently carrying the antibody. Occurrence of the disease is estimated to be doubling every year. Young people at enlistment age will be the most affected. In the past members of the Armed Forces have been six times as likely as the general population to contract venereal disease. If this trend holds for AIDS we as researchers and leaders must be very concerned about this disease and its potential for disruption of the Armed Services.

As most of you know the Department of Defense made a decision on 30 August to test all new applicants for the HTLV-III antibody. A positive test result indicates that the individual has come in contact with the AIDS virus some time in the past. It does not indicate that the individual has the virus now nor does it mean that the individual has the disease now. However, recent tests show that a very high percentage of people testing positive for the antibody do carry the virus at that time.

The testing program has already begun. It started on 10 October in the training centers and on 15 October in the MEPS. Based on this test we will deny enlistment to applicants who test positive to the antibody.

In implementing this program I wanted to maintain the current processing flow because it is critical to manning the force. We wanted to minimize the impact on shipping. Actually, as I mentioned, we don't have many straight shippers. Almost everybody we enlist goes into the Delayed Entry Program or back to Reserve or National Guard units. Only about 5% are straight shippers. They can't ship any longer. They must wait at least one day for the test results to come back. We wanted to be very accurate in the preparation of our specimens with all the traffic involved. We had some difficulty in testing initially but hopefully that has been corrected.

Applicants to the Delayed Entry Program are currently being tested at the training centers in order to allow the MEPS to continue their normal processing flow. However, by 1 October the MEPS will take over the responsibility for all AIDS testing except for active duty personnel. As you may know the plan is to begin testing the active duty force in the future.

Finally we wanted to be very careful in notifying the applicants that they were carrying the antibody. We decided to return them to the MEPS and tell them face to face as opposed to sending them a registered letter which we studied as an alternative.

To develop our plan we worked with a number of the MEPS commanders with advice from the service recruiting commands. We could have gone with a decentralized system very easily but the cost would have run to $16 or $18 per test. This high cost and the concern over quality control forced us to a national contract. One lab which would pick up nationwide and overseas and do the lab tests in one location. This would allow Walter Reed and other institutions to perform very specific quality control procedures.

We then had all of our medical Noncomissioned Officers in Charge (NCOICs) come into Chicago so we could train them very carefully. We also brought in all our medical officers - each of the 70 MEPS has one full time chief medical officer. Some of the larger ones have two. These are augmented with 400 fee based physicians. The average age of medical officers is 60. We had four over the age of 80. Just getting them to Chicago was an effort. But they really were enthusiastic about the program. The medical officers and the MEPS commanders have given the program their full support.

USMEPCOM began testing 15 October; the Navy, Air Force and Marine training centers started 1 October; and the Army training centers started 10 October. We had already completed over 17,000 tests by last Friday and we are moving forward.

How many positives are out there? It is emerging data; I don't know yet. I know the studies we did at Fort Benning showed 8 positives per 1000 for 1,400 tested. We don't know what the numbers are going to be in this program but we hope it is going to be less than 8 per 1,000.

xxx

The contract went to Damon Corporation Medical Labs in Needham, MD. The cost was $4.41 per test as opposed to the average of $16 to $18 for decentralized testing. All tests will go to Dallas by overnight express and the results electronically relayed to us each day before two o'clock. We test by doing the ELISA Screening Test twice. If the specimen is tested positive twice it is sent to Chicago for the more specific Western Blot Confirmatory Test. And if this test proves positive only then do we call the individual in to inform him or her of the results.

With the ELISA Screening Test we get the negatives back in one day. So if we pick up by noon, by noon the following day the results are back. This allows us to ship all our applicants except those who were ELISA positive. It takes three days to get the results of the Western Blot Confirmatory Tests back to the MEPS.

Issues of confidentiality are obviously a great concern for us. At the time the individual comes in for the physical exam we tell them that they are going to be tested for the HTLV-III antibody. They sign an acknowledgement form which amounts to an informed consent so it is not a complete surprise if they are called back with a positive result.

The release of the information will be a problem. We are handling it through the Office of the Secretary of Defense General Counsel. All requests are being sent there for disposition. Specific requests may go through Central Disease Control in Atlanta. We don't know at this time. We won't be indiscriminately releasing any information concerning individuals.

The notification process entails sending a letter to the individual to return to the MEPS about a medical problem. The recruiter who made the initial contact with the individual and his or her family will go to the home and pick up the applicant and return with the applicant to the MEPS. The individual will be informed of the positive results by the chief medical officer with the commander present. After notification a second blood sample will be drawn. We are doing that to be 100% positive. The individual will then be returned home by the recruiter. We will inform the individual of the results of the second test by registered mail, return receipt requested, or by telephone; probably both. If the second test comes back negative - we don't expect to have any of these - we have to research every aspect of the case to find out why.

A second letter is actually handed to the individual after the interview telling him or her that they have the antibody, but it doesn't mean they have AIDS; that they may or may not contract the disease; and that they should see their private physician to seek further advice about their health.

We also have a fact sheet developed at our request by the doctors at Walter Reed. The fact sheet is very good and is used by own people to overcome the concern they have about the disease. Put yourself in the place of the recruiter who has to pick up an applicant and then come maybe three hundred miles in six hours to

the MEPS and then return home with the individual. This is a very difficult task. The fact sheet helps calm the fears of all involved.

We have also developed a workshop program to give the recruiter training in how to handle these applicants. In developing this training we used social workers, psychiatrists, and suicide counselors. They came up with some do's and don'ts to reinforce and give confidence to the recruiters before they have to deal with this situation. We have included in this training a list of questions and answers which address possible problems the recruiter could encounter while bringing in an applicant.

We have about 7,000 recruiters and all of them will be faced with transporting a ELISA positive applicant at one time or another. We are taking this workshop to the recruiting schools where it will become part of their training.

This is the end of my prepared remarks. Thank you for your time and attention.

Rule Space: A Model for Identifying Erroneous Rules of Operation

Maurice Tatsuoka

University of Illinois at Urbana-Champaign

Introduction. It is well known that a student's total score on a test does not tell the whole story--in fact it often tells very little--about the student's achievement level and even less about the kinds of incorrect notions he or she has about the subject matter being tested. To get something close to the whole picture, we need to examine the student's response pattern, which is a vector of 1's and 0's, representing right and wrong, respectively, on the successive items, like [1,0,0,1,1,1,1,0,...,0,1,1].

However, since the vector will have as many elements as there are items on the test, it is clear that we'd be hard put to make much sense out of it unless the test is quite short. What is needed, therefore, is some way to summarize the information contained in a response-pattern vector by means of some kind of numerical index.

Suppose that an achievement test has been given to a sizable group of students, and the facility level of each item (i.e., the percentage of the group that got each item right) has been determined. Suppose, further, that the items have been rearranged from the easiest to the hardest in recording the 1's and 0's (for right and wrong) in the successive response-pattern vectors. What sort of vectors would we expect to find predominantly in the set of vectors for the whole group? From the way the items have been arranged, we should find most vectors to have more 1's toward the left and more 0's toward the right-hand part--representing a pattern of responses in which more of the easier items were passed and more of the harder ones were failed. The "ideal" pattern would be one in which all the 1's precede all the 0's. Such a vector may be called a "Guttman vector," in analogy with what are called Guttman scales in attitude scaling. Of course, Guttman vectors would be found rarely if at all in any set of real response-pattern vectors. Most of the vectors would show some 0's interspersed among the predominant 1's in the left part and some 1's among the 0's in the right-hand part.

In several vectors, we may find a rather random assortment of 1's and 0's throughout; in a few, we may even find a predominance of 0's occupying the left part (corresponding to the easier items) and more 1's in the right-hand (i.e., the harder-items) part. Both these types of response vectors would have to be regarded as anomalous or "aberrant," because the first type shows just about the same proportions of the easier and the harder items being passed, while in the second type more of the harder items are passed than the easier items. Thus an index that shows the extent to which a given response vector approaches a Guttman vector or the opposite extreme in which all the 0's precede all the 1's (a "reverse Guttman vector") could serve as a measure of the "typicality" or "atypicality" of that vector. Such an index was, in fact, developed at an early phase of our project, and it was called the "Norm Conformity Index" (NCI). However, this index has the defect that it defines "typicality" with reference to an "ideal" standard or norm that is virtually unrealized in practice.

Extended Caution Indices. The above circumstance led us (mainly Kikumi Tatsuoka) to seek a measure of typicality/atypicality that is based on a probabilistic model, namely Item Response Theory (IRT). In highly simplified

terms, what our model does is first use IRT with the one- or two-parameter logistic function to estimate the student parameter θ and the item parameter $\underline{a}$ (or parameters $\underline{a}$ and $\underline{b}$) for each student and item in a data matrix. Then, for each student, the IRT-based probability for passing each item is computed, and a matrix $(P_{ij}(\hat{\theta}_i))$ is constructed. This could be called an IRT-based theoretical data matrix, with the 0-1 entries replaced by $P_{ij}(\hat{\theta}_i)$'s. Finally, an index analogous to the NCI is computed for each student, representing the extent to which his/her observed response vector deviates from the corresponding vector of probabilities based on IRT, relative to the extent to which the group's average response vector deviates from that IRT-probability vector.

Actually, several such indices, called "Extended Caution Indices" (ECI) can be defined, depending on just what IRT-based quantities are used to replace the 1's and 0's of the Guttman vectors and their means over students. Five such ECI's were defined and discussed by K. Tatsuoka and Linn (1983).

It turned out that the one that was called ECI4, when standardized, served the best as an index for detecting anomalous response patterns in a group. (The standardization is for the purpose of making the values comparable over different θ's.) This was denoted by ζ and has been used exclusively in all our subsequent work. It also has the convenient property that, prior to being standardized, it is interpretable as a linear mapping function:

$$\begin{aligned} f(\underset{\sim}{x}) &= (\underset{\sim}{P}(\theta) - \underset{\sim}{x})'(\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta)) \\ &= \Sigma_{j=1}^{n} (P_j(\theta) - x_j)(P_j(\theta) - T(\theta)) \\ &= -(\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta))'\underset{\sim}{x} + (\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta))'\underset{\sim}{P}(\theta), \end{aligned}$$

which associates with each response pattern $\underset{\sim}{x}$ a real number $f(\underset{\sim}{x})$. Here

$$\underset{\sim}{T}(\theta)' = [T(\theta), T(\theta), \ldots, T(\theta)]'$$

is a vector whose n elements are all equal to

$$T(\theta) = (1/n)\Sigma_{j=1}^{n} P_j(\theta),$$

which is the mean, over items, of the IRT-based response probabilities for a fixed θ.

The expected value and variance of $f(\underset{\sim}{x})$ for fixed θ, denoted $f_\theta(\underset{\sim}{x})$, was shown by K. Tatsuoka (1985) to be

$$E(f_\theta(\underset{\sim}{x})|\theta) = 0$$

and

$$Var(f_\theta(\underset{\sim}{x})|\theta) = \Sigma_{j=1}^{n} P_j(\theta)Q_j(\theta)[P_j(\theta) - T(\theta)]^2$$

respectively.

Rule Space. With the standardized mapping function

$$\zeta = f_\theta(\underset{\sim}{x})/[Var(f_\theta(\underset{\sim}{x})|\theta)]^{1/2}$$

defined above, we can now map each student's response pattern $\underset{\sim}{x}$ into a point (θ, ζ) in a two-dimensional space with abscissa $\hat{\theta}$ and ordinate ζ. It was shown that the maximum likelihood estimate of θ, MLE $\hat{\theta}$, and ζ are uncorrelated. This space is called "Rule Space," because--if a student consistently uses some specific rule of operation (or algorithm) in solving all the items on a test--each rule R will yield a unique response pattern $\underset{\sim}{x}_R$ (or just $\underset{\sim}{R}$ for short). Thus the correct rule will yield the response pattern [1,1,1,...,1], and each incorrect rule will yield some specific permutation of 1's and 0's (including all 0's, of course).

Consequently, the rules that yield response patterns leading to the same MLE $\hat{\theta}$ will all be mapped into points that lie on a straight line perpendicular to the $\hat{\theta}$ axis. (Such response patterns will have the same number, Σx_j, of 1's in the case of the one-parameter logistic model, and the value of the sufficient statistic, $\Sigma a_j x_j$, will be equal in the case of the two-parameter model.)

Thus, what ζ does is to pull apart students who have the same total score (or the same sufficient statistic for θ) on a test. Students whose response patterns are typical of their group (and hence whose ζ values are small) will be represented by points close to the $\hat{\theta}$ axis, while those whose response patterns are unusual for the group (large ζ values) will be mapped into points that are above the $\hat{\theta}$ axis--the more unusual their pattern, the farther up their points will be.

But now comes the rub. The neat mapping indicated above was predicated on the assumption that each student uses some specific rule consistently throughout the test. But of course this will not be true in practice. There are bound to be some random departures from the use of a constant rule in a few of the items. Then the student's response pattern will not be mapped into a point that corresponds to consistent use of a single rule. Rather, the point will fall in the vicinity of the point for the rule that has been used for solving most of the items. These may be called "perturbations" from the "pure" rule point. Or, if we want to be more specific about the extent of perturbation--i.e., the number of items on which there was a departure from the modal rule--we may call the points "one-slip points," "two-slip points," and so on. Tatsuoka and Tatsuoka (submitted to Psychometrika) have shown that the probability of there being no more than some number s ($<n$) of slips in a test with n items is given by a compound binomial distribution. Hence, asymptotically, the perturbed points will be distributed normally around the pure rule point from which they are perturbations. Moreover, under reasonable assumptions, the variance of this distribution will--at least for rule points that are not too far from the $\hat{\theta}$ axis--be equal to the variance that was previously displayed for $f_\theta(\underset{\sim}{x})$, namely, $\Sigma_{j=1}^{n} P_j Q_j [P_j(\theta) - T(\theta)]^2$. This, along with the facts that $\hat{\theta}$ is normally distributed with mean θ and variance $1/\underset{\sim}{I}(\theta)$ and that ζ and $\hat{\theta}$ are uncorrelated, allows us to conclude that the perturbed points will follow a bivariate normal distribution with a known centroid and a known, diagonal covariance matrix. Consequently, the points corresponding to response patterns with no more than a certain number of slips will lie inside ellipses whose minor and major axes are parallel to the reference axes of rule space. These ellipses will constitute various iso-density ellipses of the bivariate normal distributions around the pure rule point. The upshot is that, if all the rule points and the observed response points that are perturbations from them were to be plotted in rule space, the result would be something like this: There would be several swarms of points, each of which would be most densely concentrated around one of the pure rule points, and would become sparser and sparser as we go farther from the center (i.e., the rule point) in any direction.

We can now explicate just what is meant by the title of this paper, "a model for identifying erroneous rules of operation," as follows: It is a technique whereby, given a student's point $(\hat{\theta}_i, \zeta_i)$ in rule space, we are enabled to decide to which one of the ellipses this point most likely belongs. This is because we have, through this model, translated our original problem into a typical problem of statistical decision theory: the problem of classifying a given point into one of several bivariate normal populations--or, more generally (as we shall soon see) into one of several 2m-variate normal populations.

One of the ways for solving such a problem is to invoke the "minimum-D^2 rule," where D^2 is the (squared) Mahalanobis generalized distance. Without loss of generality, we may assume that we have eliminated all but two rule points $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ as candidates for the rule point from which the given student's response point is a perturbation. If we denote the common covariance matrix of

the two bivariate normal distributions with $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ as their centers by $\underset{\sim}{\Sigma}$, then the D^2 of the given response point from the centroid $\underset{\sim}{R}_k$ of the Rule k ellipse is

$$D^2_{\underset{\sim}{x}k} = [(\hat{\theta}_{\underset{\sim}{x}}, \varsigma_x) - \underset{\sim}{R}_k]' \underset{\sim}{\Sigma}^{-1} [[(\hat{\theta}_{\underset{\sim}{x}}, \varsigma_{\underset{\sim}{x}}) - \underset{\sim}{R}_k] \qquad (k = 1,2)$$

Upon computing the two D^2s, we would classify $\underset{\sim}{x}$ as a perturbation from $\underset{\sim}{R}_1$ if $D^2_{\underset{\sim}{x}1} < D^2_{\underset{\sim}{x}2}$ and otherwise as a perturbation from $\underset{\sim}{R}_2$.

Once we have decided what particular rule it is that the student's response pattern is most likely to be a perturbation from (i.e., what rule he/she most likely used almost consistently except for a few "slips"), it is a short step to diagnosing the particular misconception that was most likely held by the student for him/her to have adopted that rule in the first place. This is because the rules were originally inferred from a careful error analysis of actual test papers by experienced subject-matter specialists of the material covered by the test.

Application to a Test on Subtraction of Signed Numbers. Suppose that a test consisting of the ten items listed in the first column of Table 1 (without the answers, of course) was given to a group of seventh-graders, and that four of the students each used one of the four erroneous rules described at the bottom of the table. The four pairs of columns—each pair headed by a rule number—show the answers obtained by using the four rules, respectively, and the binary score (1 or 0) for each item.

Table 1. The Binary Response Vectors for a Set of Ten Items Responded to by Four Incorrect Rules.

Items	Rule 1 Response	x_j*	Rule 2 Response	x_j	Rule 3 Response	x_j	Rule 4 Response	x_j
-3 - (-7) = +4	-4	0	+4	1	+4	1	+4	1
-2 - 8 = -10	+6	0	+6	0	+6	0	-10	1
5 - (-12) = +17	-7	0	+7	0	+17	1	-7	0
-11 - +8 = -19	-3	0	+3	0	-19	1	-19	1
9 - 4 = +5	+5	1	+5	1	+5	1	+5	1
-15 - (-9) = -6	-6	1	+6	0	-6	1	-6	1
-13 - 5 = -18	-8	0	+8	0	-8	0	-18	1
8 - (-6) = +14	+2	0	+2	0	+14	1	+2	0
-5 - +11 = -16	+6	0	+6	0	-16	1	-16	1
1 - 10 = -9	+9	0	+9	0	-9	1	-9	1

Rule 1: The student subtracts the smaller absolute values from the larger absolute value and takes the sign of the number with the larger absolute value in his/her answer.

Rule 2: The two numbers are always subtracted as seen in Rule 1 but the + sign is always taken in the answers

Rule 3: The student converts -2 - 8 and -13 - 5 into -2 + 8 and -13 + 5, respectively, but the other eight items items are converted to addition correctly. Then the right addition rule is used to answer them.

Rule 4: The student has a strange idea about the parentheses. Converts operation sign, -, to +, first. Then he/she follows the rule: if the signs of the two numbers are minus, then change the sign of the second number to a +; if the signs of the two numbers are not alike, then the sign of the second number becomes a minus.

* x_j is the score for the jth item in the binary response vector $\underset{\sim}{x}$.

We see from Table 1 that Rules 1 and 2 both yield correct answers in two of the ten problems, but not the same two. Similarly, if a student consistently uses either Rule 3 or 4, he/she would get eight problems right (again, not the same eight). Thus, consistent use of either of the first two rules or either of the second two would—assuming the discrimination parameters a_j to be equal for all of the items—respectively yield the same estimated θ values. However, examining the descriptions of the four rules at the bottom of the table, we see Rule 2 represents greater ignorance than Rule 1, and Rule 3 might be the result of careless slips while Rule 4 is quite a "weird" one. Therefore, consistent users of Rule 1 and those of Rule 2 should not be treated alike. This is probably true for the sheer purpose of guaging their achievement in signed-number subtraction, and is certainly true for the purpose of diagnosing their misconceptions and giving them remedial instruction; similar remarks hold for users of Rules 3 and 4. This is where the ECI ζ comes into play. Students who get the same θ by virtue of consistently using Rule 1 and Rule 2, respectively, would be pulled apart by their different ζ values--those using Rule 2 being plotted higher up in rule space (assuming the group to which they belong is reasonably competent). Likewise, those using Rule 4 would come higher up than Rule-3 users, because Rule 4 is a much more unusual rule. Readers who are interested in further details on how ζ works are referred to K. Tatsuoka (1983, 1985).

Sometimes, even the two-dimensional rule space will not suffice to pull the different rule points apart enough. In fact, tests of signed-number subtraction are a case in point. In such cases it may be possible to achieve better resolution, as it were, by introducing subscores for each item; e.g., a subscore for the absolute value of the answer being right or wrong, and another subscore for the sign of the answer. Then there will be one θ for each "component" and likewise one ζ for each, thus generating a rule space of four dimensions, or more generally 2m dimensions, when there are m components.

Application to Adaptive Testing. The ideas of rule space can also be used in speeding up the "convergence" in computerized adaptive tests--i.e., for getting a stable estimate of a student's θ value by administering, on the average, a much smaller number of items than in the traditional approach in which only the information value $I(\theta)$ is considered in choosing the next item. (See Lord, 1980.) In highly simplified terms, the idea is to have at hand--i.e., stored in the computer--a curve of $P_j(\theta) - T(\theta)$ plotted against $T(\theta)$ for each item in the pool. These curves will be called "Item Disparity Curves" (IDC) for subsequent reference. Then, in selecting the next item to be given to a student, we take into consideration--in addition to the $I(\theta)$--the IDC of each item remaining in the pool, and choose an item whose IDC is as far removed as possible from the IDC of the item just taken, at the current estimate $\hat{\theta}_{(k)}$ of the student's θ value.

Space does not permit indicating the proof that the method outlined above will indeed speed up the convergence to the "correct" θ value, but it is soundly based in functional analysis--which is a branch of mathematics that has many important applications in statistics and psychometrics but which, to quote Ramsay (1980), "...we in North America are handicapped by the fact that a course in [it] is seldom a part of the preparation of an applied statistician" (1982, p. 394). A proof has been made by K. Tatsuoka, and it will soon be available in a CERL Research Report of the ONR-sponsored Computerized Adaptive Testing and Measurement (CATM) Project.

References

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Ramsay, J. O. (1982). When the data are functions. Psychometrika, 47, 379-396.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. Journal of Educational Statistics, 10, 55-73.

Tatsuoka, K. K., and Linn, R. L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7, 81-96.

Tatsuoka, K. K., and Tatsuoka, M. M. Bug distribution and pattern classification. (Research Report 85-3-ONR). Urbana: University of Illinois, Computer-Based Education Research Laboratory. (Submitted to Psychometrika)

Appendix

Proof that $f(\underset{\sim}{x})$ Increases with Unusualness of Response Pattern $\underset{\sim}{x}$. It was stated without proof, on the third page of this paper, that the standardized version, ζ, of $f(\underset{\sim}{x})$ becomes larger as the response pattern $\underset{\sim}{x}$ becomes more "unusual" for the group of which the responder is a member. We now prove this. Expanding the second expression for $f(\underset{\sim}{x})$ given on page two, we get

$$f(\underset{\sim}{x}) = \Sigma_{j=1}^{n} P_j(\theta)[P_j(\theta) - T(\theta)] - \Sigma_{j=1}^{n} x_j[P_j(\theta) - T(\theta)].$$

We may assume, without loss of generality, that the items have been numbered in ascending order of difficulty so that, for each θ, we have

$$P_1(\theta) > P_2(\theta) > \ldots P_n(\theta) .$$

(The order may differ for different values of θ.) Then, since $T(\theta)$ is the mean of the n $P_j(\theta)$ values, there must be some k such that

$$P_j(\theta) - T(\theta) > 0 \qquad \text{for all} \quad j \leq k$$

and

$$P_j(\theta) - T(\theta) < 0 \qquad \text{for all} \quad j > k .$$

This, together with the fact that the first sum in the above expression for $f(\underset{\sim}{x})$ is a constant for any fixed θ, implies that for any response pattern that has a small number of $x_j = 1$ for $j \leq k$ (i.e., the easier items) and a large number of $x_j = 1$ for $j > k$ (the harder items), the quantity $f(\underset{\sim}{x})$ will have a larger numerical value than for a response pattern for which the opposite is true. But the first-mentioned type of response pattern (with a small number of corrects among the easier items and a large number among the harder items) is clearly an "unusual" or atypical one. This shows that, for a fixed θ, the larger $f(\underset{\sim}{x})$ is, the more unusual is the response pattern $\underset{\sim}{x}$--for the group in which the items were calibrated. The standardization to get ζ makes this property hold across different values of θ.

Acknowledgement

This research was sponsored by the Personnel and Training Research Program, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-82-K-0604. Contract Authority Identification Number NR 150-495.

Implementation of a Computer System to Support Diagnostic Testing

C. David Vale
Assessment Systems Corporation
St. Paul, Minnesota

The MicroCAT™ Testing System, developed by Assessment Systems Corporation, is a complete system of computer programs to support computerized adaptive testing (CAT). MicroCAT runs on an IBM Personal Computer and includes facilities for performing all of the functions necessary to implement adaptive tests. These functions include authoring test items, calibrating items according to item response theory (IRT) models, authoring tests using any of a variety of CAT strategies, and administering tests. The details of the MicroCAT system have been described elsewhere (Assessment Systems Corporation, 1984; Vale, in preparation, a, b) and will not be detailed here.

MicroCAT served as the basis for the testing system implemented in the Basic Electricity and Electronics (BE&E) School of the Naval Training Center (NTC) in San Diego. The purpose of this implementation was twofold: to provide a delivery system for the diagnostic testing techniques developed by the University of Illinois, and to provide an environment in which to evaluate the MicroCAT system, which was not then in commercial release.

There were two general requirements that the system for NTC had to meet. First, it had to administer the current NTC tests in a form that was comparable to their original mode of administration. (This was to allow the collection of data on the current forms in an operational testing environment.) Second, the system had to be capable of administering tests according to the prescribed diagnostic strategies.

In the standard mode of administration at NTC, a student is assigned a microfiche card which contains a test. He or she then reports to a carrel with a microfiche reader and responds to the test questions on an optically scannable answer sheet. After completing the test, the student puts the answer sheet into an optical scanner that is connected to a computer terminal which, in turn, is connected to a mainframe computer in Memphis, Tennessee. A computer-managed instruction program running on that computer scores the results, updates the student's records in the database, and reports the score to the student. It also tells the student which test to take next.

Our goal in designing the MicroCAT interface to this mode of testing was to be able to install the computerized testing system in such a manner that it would be perceived by examinees as comparable to the microfiche version of the test. Its operation also had to be completely transparent to MIISA, the computer-managed instruction program, so that the results of the computerized tests could be used operationally in the computerized instructional management system. MIISA is a very complex program that manages all of the instructional assignments and record keeping at NTC. It was written several years ago, runs on a mainframe computer, and is very resistant to change of any kind. Any approach to implementation that required reprogramming of MIISA was not viable.

Although the MicroCAT system was originally designed for stand-alone testing, a general networking capability was added to allow a proctor to assign and monitor tests at several stations from a single proctor's station. This standard MicroCAT networking system was still unable to communicate with MIISA, but the proctoring station provided a good starting point for the connection. To achieve the transparency desired, the proctoring program was adapted so that it could interact with MIISA as well as with a proctor. It was changed to make the proctoring terminal emulate one of the terminals that normally communicated with the MIISA mainframe. To accomplish this, we made the proctoring station's serial port look like a GE Terminet printing terminal connected to an optical scanner. As far as MIISA was concerned, it was communicating with an optical scanner connected to a terminal. When an examinee reported to the testing room, the proctor assigned him or her to a testing station, where the examinee entered his or her social security number. The testing station communicated this to the proctoring station, which queried MIISA for the appropriate test to administer and assigned it to the examinee's station. After the examinee completed the test, the proctoring station passed the responses to MIISA, obtained the scores, and printed them out for the examinee on the system printer. From MIISA's perspective, the computerized system was capable of performing all of the functions usually done on paper.

Figure 1 diagrams the testing system as implemented at NTC. It contained eighteen terminals. Fifteen of these were testing stations, two were network servers (which contained all of the tests and response data), and one was a proctoring station. One network server would have been sufficient; the second one was for backup in case the first one failed. One of the testing stations also had the hardware necessary to convert it to a proctoring station if the first one failed.

There was a fear among the Navy Chiefs in charge of testing that it might be difficult to convince the examinees that the computerized mode was equivalent to the paper-and-pencil mode. Their concern was that if tests were administered in two forms, superstitions would develop among the students regarding which mode was easier. Obviously, this problem would be exacerbated if there was any substance to the claim.

Specifically, three factors were initially assumed to be related to the examinees' acceptance of the computerized mode as an equivalent mode of testing: (1) the system had to respond quickly with the next item after the examinee answered the previous one, (2) it could not lose an examinee's work if any portion of the system failed, and (3) it had to support standard test-taking strategies such as skipping difficult items and coming back to them later.

MicroCAT was fast enough to satisfy the first requirement, but originally it did not support the other two features. Therefore, a recovery feature was added to save the examinee's responses on a diskette as soon as they were made. If a testing station failed, the examinee could simply remove the diskette in his or her station, put it in another station, enter his or her social security number, and continue the test with no loss of data.

The ability to skip items and later return to them was also added to the MicroCAT system. At the end of the test, the examinee is asked if he or she would like to review all of the items, some specific ones, or just those he or she skipped. A

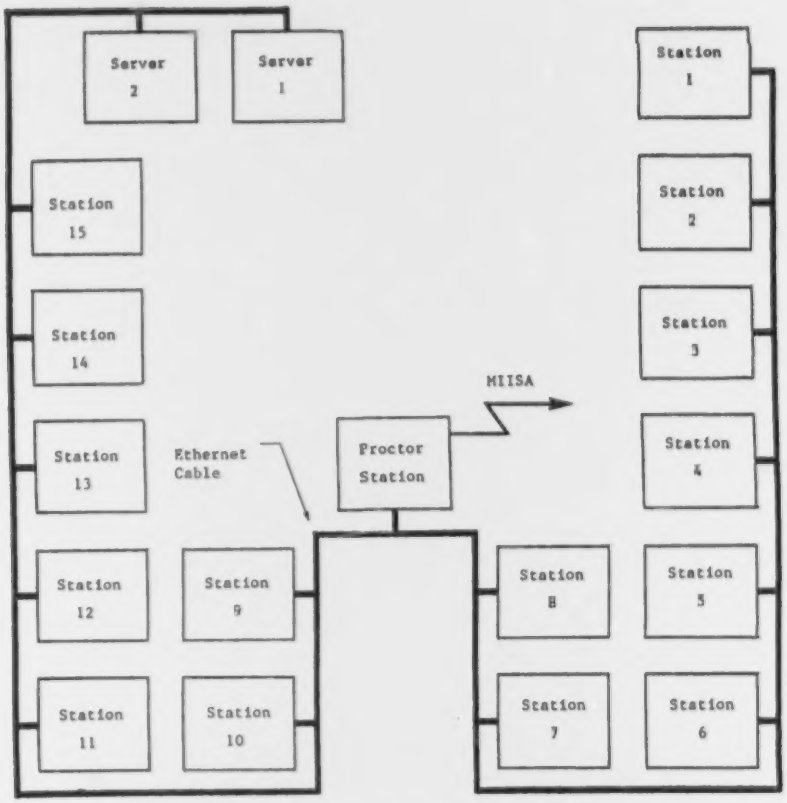


Figure 1. Structure of the NTC Implementation

review is granted as desired, and the specified items are re-presented to the examinee with his or her previous response shown. The examinee can then change the response or leave it as it was.

After these features were added to the MicroCAT system, it was installed for a small-scale evaluation. In general, the initial system ran without any serious errors. Several small problems did require attention, however. First, as the system was originally set up, it was possible for students to take two tests simultaneously. Since they received no feedback regarding the correctness of responses, there was really no advantage to be gained by doing this. Nevertheless, the problem was corrected by modifying the proctoring program to keep track of who was on the system and to allow examinees to take only one test at a time. The second potential problem was that hitting some combinations of keys could abort the testing system. This was solved by disabling all of these combinations except one that required the examinee to simultaneously push three keys. It is highly unlikely that anyone would accidentally hit all three keys at once; anyone who did hit them would probably be deliberately trying to abort the system. It was not possible or worthwhile to subvert such an attempt, because a determined examinee could always reset the station simply by unplugging it.

The system was implemented operationally in August of 1984. The test items presented were similar in format to the one shown in Figure 2 (which is not an active test item). Most examinees had no trouble with the new system. The major surprise on the first day of testing was that, on the average, examinees took ten minutes to respond to an item. This did not begin to tax the capacity of the network, which was designed to handle a response from each examinee every ten seconds.

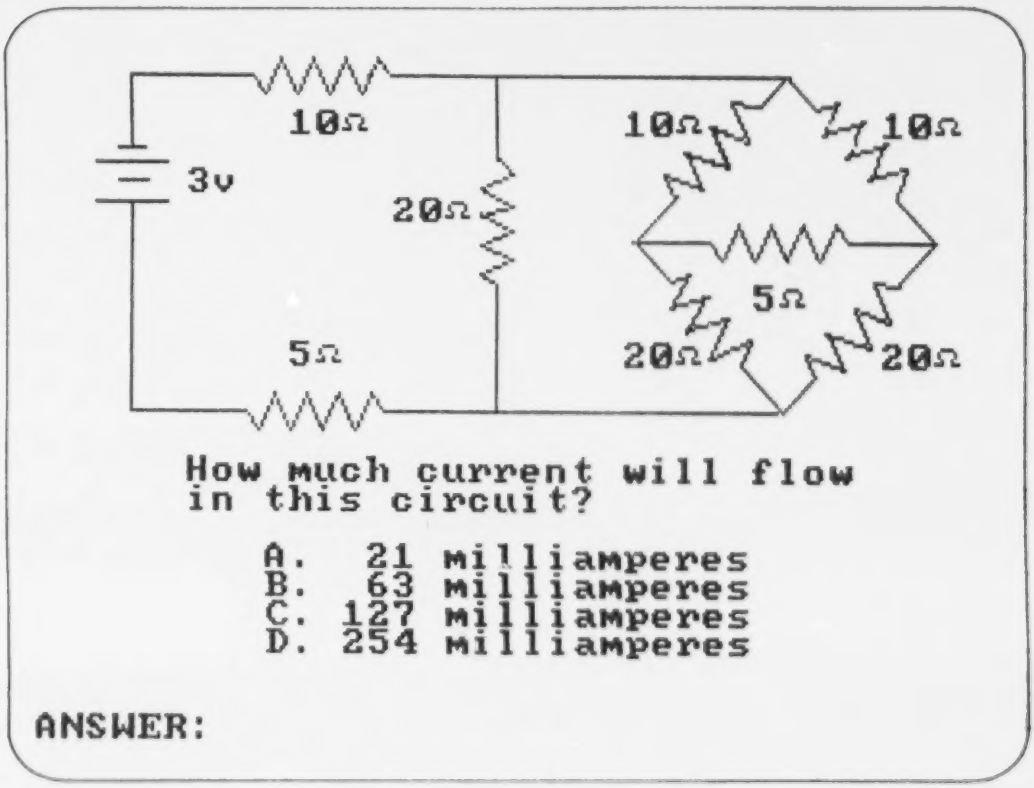


Figure 2. Sample Electronics Item

Since the data collected by computer administration were to be analyzed by the University of Illinois, a data transfer scheme was needed. The MIISA link is a real-time link in that testing waits for communication. Transferring the data to the University of Illinois, on the other hand, had to be done only when the data were needed or when the disks on the NTC network were full. A system was developed whereby the test proctor periodically dumped the data from the system disks to two sets of diskettes, one for the University of Illinois and one for backup. After dumping the data, the proctor's instructions were to mail one set to the University of Illinois and to keep the backup set until receipt was confirmed. The data on the system disk were erased after the diskettes were made. Except for the difficulty of getting the proctor to make the data diskettes on a regular basis, this scheme worked well.

Testing has not been interrupted because of any system problems. It was interrupted for several weeks, however, by the implementation of new versions of the tests. The frequent changes in tests, which had not been anticipated when the system was installed, required frequent communication with the University of Illinois. It had been intended that the University of Illinois would do the test development and then either manually install the tests in the San Diego system or mail complete test files with installation programs to be run by the proctor. However, as the test changes became more frequent, it became apparent that it would be more efficient for NTC personnel to make the changes themselves and install the tests.

Test development in the MicroCAT system is a three-stage process. First, the items are authored using the system's Graphics Item Banker. Then the test is specified using an authoring language. Finally, the authoring language is compiled, a process that reformats the items and processes the instructions in a manner that allows items to be presented rapidly. Implementing a test in the NTC system required the further step of copying the compiled test onto the appropriate disk volume.

NTC test administration personnel mastered the process with relative ease. However, several problems arose that had bothered us in the early part of this effort but which we had forgotten until the NTC personnel began to use the system. One such problem was that if diskettes were swapped while the item banker was running, a bank would be destroyed. We originally circumvented this problem by not swapping diskettes, but this solution was obviously not optimal. We therefore wrote a utility program that could recover a bank destroyed in this manner. The other problem that we re-discovered was that, using the Ethernet network from 3Com, two people sharing a disk volume can, under certain circumstances, destroy each other's work. For example, NTC personnel recently destroyed an item bank by writing portions of a memo over it. Fortunately, the new program was able to restore most of what was lost.

In general, the implementation at NTC has been successful. The MicroCAT system has flawlessly performed most of the tasks required of it. More than 2,400 items have been banked for this application, approximately 50 different tests have been implemented, and over 1,500 tests have been administered. Informal evidence from the BE&E School suggests that the system is fast enough for administering all of the tests and that the tests are perceived as psychologically parallel to the microfiche form (although we have heard that the computer display is easier to read than the microfiche). The system has not yet been used for diagnostic testing, but custom interfaces (portions of the program that allow programmers to augment the MicroCAT system) have been provided to allow diagnostic testing routines to be implemented within the MicroCAT system.

The MicroCAT Testing System is now a commercial product and the contract that supported its development and implementation at NTC is near an end. Its use at NTC will continue and the University of Illinois will implement the diagnostic strategies in the near future. It has been a successful implementation that has demonstrated the relative ease with which the transition can be made from paper-and-pencil testing methods to computerized testing, even when the new system must be integrated into a complex instructional management system that is already in place.

References

Assessment Systems Corporation. (1984). *User's manual for the MicroCAT testing system.* St. Paul: Author.

Vale, C. D. (in preparation, a). *Implementation of a microcomputer-based testing system in a Navy training environment* (Research Report ONR-85-XX). St. Paul: Assessment Systems Corporation.

Vale, C. D. (in preparation, b). *MCATL: A language for authoring computerized adaptive tests* (Research Report ONR-85-XX). St. Paul: Assessment Systems Corporation.

Acknowledgments

The preparation of this paper and part of the development of the MicroCAT system was supported by the Personnel Research and Training Programs, Psychological Sciences Division, Office of Naval Research, under contract N00014-83-C-0634 (NR 507-002).

MicroCAT is a trademark of Assessment Systems Corporation.

APPLICATION OF RULE SPACE IN A NAVY TESTING ENVIRONMENT

by John M. Eddins
University of Illinois at Urbana-Champaign

Introduction

The MicroCAT testing system was installed in the BE&E School at the San Diego Naval Training Center in order to extend to an actual training environment the theoretical work in diagnostic adaptive testing described by Dr. Tatsuoka, and reported in detail in Tatsuoka (1985, in press), Tatsuoka & Tatsuoka (1985) and Tatsuoka, Tatsuoka & Baillie (1985). I will present first a summary of our overall plans for the project, next our progress to date, including some of the problems we have encountered, and finally what we see as the next steps.

Summary of Project Plans

The principal goal of the project is to develop one or more diagnostic adaptive tests for the BE&E curriculum using the rule space model, and to evaluate the effectiveness of these tests with data collected from pilot groups of Navy trainees. The overall plan can be outlined in five phases.

1. Collect and analyze data from current Navy tests.
2. Link rule space procedures to the MicroCAT system.
3. Add experimental items to current tests.
4. Create experimental tests.
5. Give experimental tests to pilot groups.
6. Report the results.

Substantial progress has been made on the first two phases, and we are currently working on the third.

Progress to Date

1. Data collection and analysis

The MicroCAT system was installed and tested during the summer of 1984, but ongoing revisions of the instructional materials and tests delayed full implementation until spring, 1985.

BE&E trainees at the San Diego base began testing with the system in late March, 1985, and data was collected for Modules 1, 2, 4, 5, 6 and 7 during April through August, 1985. As it turned out, one more revision of the tests was necessary. The revised tests are now on line and the computer testing lab is back in operation.

As each BE&E trainee takes a test, the MicroCAT system stores a record which includes, for each item, the item number, answer key, student's response, and time to respond. These data are added to a file which eventually is read to a diskette and mailed to us at the University of Illinois. Our plan is to gather enough data to analyze the items

statistically and estimate item parameters, on the assumption that some existing items can serve as a starting point for developing items applicable to rule space. This requires a minimum of 200 to 300 subjects, with a significant percentage giving wrong answers. Between April and August we collected data for about 1500 subjects; however, these were divided among five versions each of six tests, leaving only around fifty for each test version. The different test versions for Mod 4 proved to be identical except for the ordering of the answer foils, so we re-ordered the data appropriately and merged it into a composite set, giving us 235 subjects. An analysis of variance across the different forms confirmed their equivalence.

A summary of responses for the items in this dataset is shown in Table 1. A high percentage of the responses on most items are correct, and several items show essentially no discrimination. These data are based on the first pass through the test, and they include partial tests taken for remediation, hence the substantial number of skipped items.

Table 1.

BE&E Test, Mod. 4 -- Summary of Responses.

All test forms equivalent to form 1.

Item#	Key	Rsp1	Rsp2	Rsp3	Rsp4	Skip
1	1	196	8	1	1	29
2	4	9	25	0	165	36
3	3	2	40	125	24	44
4	1	169	7	19	6	34
5	1	207	0	1	0	27
6	3	1	12	189	0	33
7	3	0	0	202	1	32
8	1	196	2	2	0	35
9	4	1	0	2	199	33
10	3	9	2	188	0	36
11	1	204	0	9	0	22
12	2	16	178	6	1	34
13	3	19	41	109	27	39
14	3	0	0	212	1	22
15	1	152	24	11	20	28
16	4	36	14	0	154	31
17	4	10	18	45	101	61
18	4	2	13	2	168	50
19	3	0	6	139	8	82
20	3	3	14	168	2	48
21	3	8	1	177	0	49
22	2	2	197	1	0	35
23	3	2	2	187	1	43
24	2	17	120	0	12	86
25	4	19	5	1	166	44
26	2	4	197	15	*	19
27	1	151	21	42	*	21
28	3	2	26	186	*	21
29	1	172	16	20	*	27
30	2	8	202	6	*	19
31	1	205	1	6	*	23
32	1	152	56	5	*	22
33	1	70	120	23	*	22
34	3	24	14	175	*	22
35	2	17	187	12	*	19

*Only three choices for these items.

Our attempts to estimate item parameters were frustrated at first because of the limited number of both items and subjects, and because of the large number of skipped items resulting from partial tests taken as remedials. We eliminated the partial tests, leaving 193 subjects, selected 22 of the most promising items and succeeded in estimating the A and B parameters for these items. We used a computer program created by Yamamoto, Baillie and Tatsuoka which implements an EM algorithm developed by Bock and Aitkin (1981). The EM method has the advantage of being able to estimate item parameters with relatively few items. Results are shown in Table 2.

Table 2.

BE&E Test, Mod. 4 -- Item Parameters for Selected Items.

Item#	A (discrimination)	B (difficulty)
10	.92832	-.49033
11	.80779	-.76712
12	.48945	-.84937
13	.80148	1.00827
15	.29399	-.72174
16	.55109	.09081
18	2.42278	.65190
19	2.33766	.91749
20	2.35894	.63324
21	3.77397	.64296
23	2.87624	.39554
25	2.93106	.71003
26	.72511	-.66316
27	.86399	.52368
28	1.07095	.07809
29	.77854	.10053
30	.38718	-2.75522
31	.98740	-.53114
32	.38085	-.18672
33	1.16947	1.68272
34	.88895	.15271
35	.62432	-.52315

2. Linking of Rule Space Procedures with MicroCAT

To set up a test to be administered on the MicroCAT system, a bank of test items is provided, together with a file which lists the items to be administered and specifies the logical basis on which each successive item is to be chosen. The choice can range from a simple top down sequence of all items on the list to a decision computed by a program external to the MicroCAT system and passed back to it. In our case, administering a diagnostic adaptive test using rule space is in the latter category.

At the beginning of a test, item parameters stored in the item bank are read into an array in memory (Figure 1). As each test item is administered the student's response is stored, and program control is passed to the rule space programs along with student response information.

The student's response is scored as right or wrong (1 or 0), the weighted distances to all remaining items are calculated (Tatsuoka & Tatsuoka, 1985), and the next item is selected for optimal speed of convergence. If sufficient convergence has been achieved, then the error probabilities are estimated between the current point and the centroids of the ellipses stored in the bug information bank. If the minimum value of these probabilities satisfies a specified criterion, next item is set for stop; if not, next item is not changed. Next item and program control are then returned to the MicroCAT test driver. The computer programs for the rule space procedures were developed by Robert Baillie. Details of the theoretical basis for these procedures will be described in a future report.

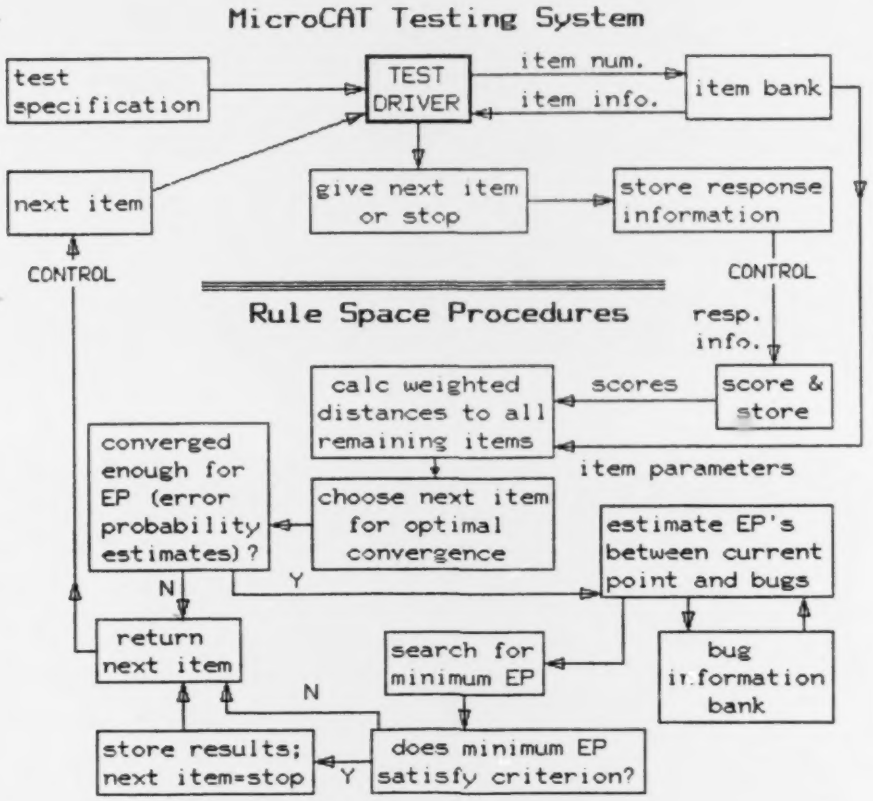


Figure 1. Linking of Rule Space procedures with MicroCAT.

Future Plans

The principle hurdle remaining before we can actually create and install experimental test items is a detailed analysis and verification of specific learning tasks to be addressed, together with hypotheses as to causes and results of various types of errors. Designing a diagnostic test for use with the rule space procedures requires that the test items be constructed with specific characteristics, and these item characteristics must be analyzed and verified with extensive data collection and computer analysis. Parallel versions of the items also are needed.

At present we are examining in detail the concepts, operations and errors represented by the Navy's BE&E tests. Three items selected from the Mod 4 test will illustrate something of the nature and complexity of the task. In the first item (A), a simple series circuit containing five resistors is shown and the student is asked the effect on voltage drops at an open resistor -- increase (1), decrease (2) or no change (3). The second item (B) shows the same circuit and answer choices, but asks the effect on current at a shorted resistor. The third item (C) shows a series circuit with values given for applied voltage and two of the resistors, one of which is open. The student is asked what a voltmeter across the open will indicate.

For item A, 56% of the students chose (2) -- decrease in voltage drop at an open; for item B, 26% chose (2) -- decrease in current at a short; for item C, 18% chose (1) -- zero volts read across an open. Out of 52 students choosing (2) for item B, 34 also chose (2) for item A; out of 36 choosing (1) for item C, 27 also chose (2) for item A; but out of the total group only 5 chose all three. Apparently, items B(2) and C(1) represent different errors, while item A(2) subsumes both errors. This is also confirmed intuitively. Response (2) to item B probably results from confusing shorts with opens (the result of a short often is an open). Response (1) to item C probably represents confusion about the nature of voltage drop (nothing "happens" across a break, so voltage drop = 0). These are different misconceptions, yet response (2) to item A can result from either one.

The task before us is to pursue this type of inquiry until we have isolated and charted a sufficient number of such relationships among items to enable us to describe several common errors and specify test items accordingly. We also are investigating ways to utilize the computer to speed up and simplify this otherwise tedious and complex task. What we are looking at here is the relationship among just one foil for each of three items. To systematically consider such relationships among all foils for all items obviously is a task for a computer. The critical human task is defining the relevant concepts and errors.

With a multiple choice test it is important that each foil be designed to provide specific information, so that the choice of the next item will be based on more than simply a right-wrong score. Our first experimental test items will likely be items taken from the current tests with some changes in the foils or the circuit values. We have the facility to insert these items into the existing tests without disturbing the regular scoring. We could, for instance, insert five extra items at the end of the Mod 4 tests which never would appear in the Navy's computer records but would provide us with response data.

As a simple illustration, assume that an item such as A is found to suggest four possible errors. We present this item first, followed by several others which confirm some errors and eliminate others. This can be diagrammed in a process network where each node represents a decision branch. By tracking these decisions the computer can identify the student's errors, provided the student is entirely consistent and is following a predicted route. Unfortunately, the realities of human behavior do not conform to such a deterministic model. By using pattern classification techniques, the rule space approach adds an element of probability to the model, with the promise of much better rates of error detection.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443-459.

Tatsuoka, K. K. (in press). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. Behaviormetrika.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classifcation approach. Journal of Educational Statistics, 10, 1, 55-73.

Tatsuoka, K. K., & Tatsuoka, M. M. (1985). Bug distribution and pattern classification (Technical Report 85-3-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K. K., Tatsuoka, M. M., & Baillie, R. (1984). Application of adaptive testing to a fraction test (Technical Report 84-3-NIE). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

An Overview of the Accelerated CAT-ASVAB Program

by

W. A. Sands

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

Group-Administered Aptitude Tests

Group-administered aptitude tests can be described as conventional, paper-and-pencil tests and Computerized Adaptive Tests (CAT). As the names indicate, these differ in administration mode. Less obviously, but perhaps more importantly, they also differ in the way in which items are selected for administration. In the usual paper-and-pencil test, all examinees are administered the same items in the same sequence. In contrast, a CAT instrument is dynamically tailored to the measured ability level of the individual examinee, during the course of the test administration. This means that, at least potentially, every individual receives a different test.

Typically, in a CAT administration, the first item selected for administration is one of medium difficulty, since we know nothing about the examinee's ability level. If the examinee responds correctly, the ability estimate is raised to above average, and a more difficult item is selected for administration. If the examinee answers this second item incorrectly, the ability estimate is lowered somewhat through the updating procedure. As a result, an easier item is selected as the third question. This process of selecting an item, scoring the examinee's response, updating the ability estimate, and choosing the next item for administration continues until some stopping rule is reached. This test termination criterion may be either the administration of a prespecified number of items (fixed length testing), or the administration of items until the ability estimate meets a prespecified level of precision (variable length testing).

Armed Services Vocational Aptitude Battery (ASVAB)

The ASVAB is a conventionally-administered, paper-and-pencil aptitude test battery used by all the U.S. military services for both enlistment eligibility screening and for subsequent classification and placement into entry-level training. The paper-and-pencil version of the battery (P&P-ASVAB) includes eight power tests and two speeded tests. Administration time for P&P-ASVAB takes about three and one-half hours.

The P&P-ASVAB is administered under two large-scale testing programs. The Production Testing Program involves the administration of the battery in the 68

Military Entrance Processing Stations (MEPS) and in about 900 Mobile Examining Team (MET) sites located across the country. The Student Testing Program is administered in about 14,000 high schools. These two testing programs are quite large, each involving the administration of the battery to between 800,000 and 1,000,000 persons annually.

COMPUTERIZED ADAPTIVE TESTING VERSION OF ASVAB (CAT-ASVAB)

Objectives

The Computerized Adaptive Testing (CAT-ASVAB) Program has two broad objectives. The first involves the development of a system that automates the administration, test scoring, and computation of the Armed Forces Qualification Test (AFQT) score and various other composite scores derived from ASVAB used by the individual military services. Such a system must be capable of use in both the fixed-base MEPS and in the portable testing environment of the MET sites, while interfacing with the existing score reporting system. The second objective of the CAT-ASVAB program is to evaluate the suitability of CAT-ASVAB as replacement for the P&P-ASVAB in the Production Testing Program.

Approach

The original approach to the development of the CAT-ASVAB System was a three-stage competitive "flyoff" between three contractors from private industry. During the first stage, the three contractors developed system design concepts and supporting analyses. In the second stage, limited production models were to be developed, field-tested, and evaluated. The final stage would have involved one of the three original contractors going into full-scale production, deployment, and implementation.

The approach has changed as a result of three factors. First, the timelines submitted by the contractors for Stage 2 were considerably longer than we had planned. Secondly, some remarkable advances have been made in microcomputer technology during the past few years. Finally, LTGEN E. A. Chavarrie, Deputy Assistant Secretary for Military, Manpower and Personnel Policy, in his keynote address at the last MTA convention in Munich, provided strong encouragement for reducing the long timelines, commensurate with meeting the performance objectives of the program. As a result of these influences, we have adopted a markedly different approach. With a focus on early implementation, we have initiated work on the Accelerated CAT-ASVAB Program (ACAP).

ACCELERATED CAT-ASVAB PROGRAM (ACAP)

Objective

The objective of ACAP is to field-test CAT-ASVAB as soon as possible. In pursuit of this objective, we will procure off-the-shelf, commercially-available microcomputer equipment. Software will be designed and developed in-house at NPRDC.

Evaluation Criteria

ACAP will be designed to meet the nine evaluation criteria originally established for the full-scale version of CAT-ASVAB: (1) performance, (2) suitability, (3) reliability, (4) maintainability, (5) ease of use, (6) security, (7) affordability, (8) flexibility/expandability, and (9) psychometric acceptability. Each of these nine major criteria includes numerous subcriteria.

The performance criterion includes both general and specific requirements. The system must automate the current P&P-ASVAB functions and anticipated additional functions of CAT-ASVAB. System response time cannot exceed a maximum of two seconds, and this response time must be independent of the number of examinees taking the test (system load). The display must have a resolution of 400x300 pixels, and the test must be displayed in 7x9 characters. The system must support an interface with the existing MEPS Reporting System, and the CAT-ASVAB Maintenance and Psychometric (CAMP) facility. The computer software should employ a "top-down", structured design, use a high-level language, and be adequately documented.

Suitability requires the system to operate in a normal office environment in terms of temperature and humidity. No significant modification to existing facilities should be required (e.g., electrical power). There should be no necessity for specially skilled operators and no significant staffing changes should be required. Finally, the system must be portable to support testing in the MET site environment.

Reliability is an important concern for the system. It is imperative that the system be available and operate reliably for scheduled testing sessions. The system must be capable of restarting from the point of failure and recovering from failure without loss of data.

Maintainability is a major consideration for any large scale computer system. No skilled technicians are to be required. The hardware/software must incorporate self-diagnostic capabilities which can be readily understood by test administrators. An adequate integrated logistics support system must be established and maintained for the life of the system.

The system must be easy to use, both for the test administrator and the examinee. No computer experience or expertise should be required. Set-up procedures should be clear, unambiguous, and adequately documented. The display legibility and resolution should support both text and graphics material. Introduction of experimental test items should be transparent to both the examinee and the test administrator.

Security is an important system consideration. The item sequence should be unpredictable. Measures must be taken to prevent printout or inspection of the item files. Use of the system must be limited to authorized personnel. A multi-level password access procedure should be implemented. System access must produce an audit trail which can be inspected by system managers. Finally, the system should be designed to minimize equipment theft, by reducing or eliminating pilferable components.

The system must be affordable, i.e., the life-cycle cost of CAT-ASVAB must be comparable to that of P&P-ASVAB. At present, an economic analysis of the system is being conducted under contract.

Flexibility/expandability is an important dimension of the system. In addition to supporting the delivery of a CAT version of ASVAB, the system must allow for future, add-on peripheral devices. While present plans call for the examinee to use a specially-designed keypad input device, the system must support standard keyboard input. A programmable, high-precision clock is essential, as future testing will almost certainly involve the measurement of response latency. Other future testing possibilities include provision to measure a person's ability to identify and track a moving target. This will require a system capable of internal and/or external expansion and provisions for additional interfaces (e.g., a joystick).

Finally, the system must be psychometrically acceptable. CAT-ASVAB must measure the same aptitudes measured by P&P-ASVAB. The CAT-ASVAB and P&P-ASVAB versions of the battery must be equated to insure that the scores are interchangeable. The CAT-ASVAB system must meet stringent professional test standards.

Progress/Plans

We have already achieved several goals. We have developed the functional requirements for the system. An equating plan has been developed, and is currently under review by policy and technical representatives from each of the services and the U.S. Military Entrance Processing Command (USMEPCOM). We have procured a small number of development systems to begin software design/development. A procurement action to obtain equating systems has been initiated and is in progress.

Plans include completion of the design and development of the ACAP software in 1986. Data collection and extensive statistical analyses will take about a year. We intend to begin the initial operational test and evaluation in selected MEPS (and their satellite MET sites) in 1987.

CONCLUSION

There have been a number of important changes in the development of the CAT-ASVAB system since the last Military Testing Association meeting in November 1984. Emphasizing the importance of demonstrating the extensive capabilities of an adaptive testing system for ASVAB, we are concentrating on the Accelerated CAT-ASVAB Program (ACAP). However, ACAP is viewed as an interim system, not as a replacement for the full-scale development of CAT-ASVAB. We plan to use "lessons learned" from the limited deployment of ACAP to strengthen our functional requirements specifications for the development of the full-scale CAT-ASVAB system.

EVALUATION OF THREE LOCAL CAT-ASVAB NETWORK DESIGNS

Ronald B. Tiggle, Ph.D.
and
Mr. Bernard A. Rafacz
Navy Personnel Research and Development Center
San Diego, California, 92152-6800

NAVPERSRANDCEN is currently involved in a major system development effort that is concerned with the research, development, and eventual implementation of a Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery(CAT-ASVAB). The goal of this effort is the implementation of CAT-ASVAB on a nationwide distributed computer network. This network will permit the United States Military Entrance Processing Command (USMEPCOM) to adaptively administer the ASVAB to civilian applicants for military service. The CAT-ASVAB System is intended to replace the operational paper-and-pencil battery (P&P-ASVAB) currently used for selection and classification of enlisted personnel.

P&P-ASVAB testing currently occurs at 68 Military Entrance Processing Stations (MEPS), two substations and approximately 900 field locations identified as Mobile Examining Team (MET) sites. The MEPS/MET sites are under the administrative responsibility of USMEPCOM. CAT- ASVAB System components will be used in both of the MEPS and MET testing sites. The system must provide an automated, on-line system for test delivery and score reporting using adaptive, conventional, and timed psychometric tests. Item response theory (Lord, 1980) constitutes the theoretical foundation for CAT-ASVAB adaptive testing.

CAT-ASVAB System Concept

In order to propose a system concept/design for a Local CAT- ASVAB Network (LCN), it is necessary to have available a set of specifications and standards for the performance of the system once implemented. Fortunately a government-specified set of performance standards exists, and it has been documented in the CAT-ASVAB Stage 2 Full Scale Development (FSD) Request for Proposal. This document outlines the functional requirements for the development of computer hardware specific to CAT-ASVAB functions. Before the FSD is implemented, an early, smaller-scale development known as the Accelerated CAT-ASVAB Project(ACAP) will be completed to provide pilot data in support of the CAT Stage 2 FSD. The scope of the ACAP effort will conform as much as possible to the CAT-ASVAB Stage 2 FSD. However, the CAT- ASVAB functions to be addressed under ACAP will be dependent on the computer hardware design that is selected. Unlike the FSD, it is not a goal of ACAP to develop computer hardware specific to CAT- ASVAB functions. Rather a system design will be selected from a set of candidate designs, and then commercially-available computer systems will be surveyed in order to identify the most appropriate system to meet the functional requirements of CAT-ASVAB.

CAT-ASVAB Operational Requirement

The primary requirement of the CAT-ASVAB System is that it be capable of administering a battery of instruments, equivalent to the present components of the P&P-ASVAB. The current production battery (P&P- ASVAB) includes 10 tests; eight of these are cognitive power tests; two are speeded tests. However, once it is implemented, CAT-ASVAB will be capable of administering other cognitive and non-cognitive operational and experimental instruments, as determined by DoD policy.

At present, 20 percent of the production P&P-ASVAB testing occurs at MEPS; the remaining 80 percent occurs at MET test sites.

MEPS P&P-ASVAB testing is currently but one part of the processing of applicants for enlistment, and it occurs at fixed-site locations in relatively controlled environments. Experienced and well-trained examiners conduct the testing sessions. In contrast to the MEPS testing environment, the MET site testing is, in a large number of cases, administered by an Office of Personnel Management (OPM) employee working under a service agreement with DoD. For the most part, MET site testing is conducted in borrowed facilities on an ad-hoc basis. USMEPCOM has no permanent control over the MET facilities, and no authority to modify them. Thus, the CAT-ASVAB System will be required to be used by non-USMEPCOM examiners who must set-up and take- down, possibly even transport, the CAT-ASVAB testing equipment to and from non-USMEPCOM facilities as required to support examining schedules.

Functional Requirements

Based on the functional specifications stated in the CAT Stage 2 RFP, it is intended that a Local CAT Network(LCN) be developed that would permit the administration of CAT-ASVAB to civilian applicants at any of the MEPS or MET sites within CONUS. An LCN would consists of (up to 24) Examinee Test (ET) Stations linked to a single Test Administration (TA) Station via a hard-wired electronic telecommunications line. In addition, a Data Handling Computer (DHC) would reside at each MEPS to support the telecommunications function among LCN units located at the MEPS and the MET sites for that MEPS, and with USMEPCOM MEPS computer systems to be used for archiving CAT-ASVAB data.

MEPS Sites. The functional capability that is required at the MEPS, in terms of psychometric testing, is identical to that required at the METS. Specifically, MEPS equipment is stationary, but identical to MET site equipment. Identical LCN components at MEPS and MET sites are necessary to accommodate the equating of CAT-ASVAB, commonality of equipment for software and hardware maintenance purposes, and to permit the cost effective sharing of equipment across both types of sites. In contrast to most MET sites, each examiner at a MEPS testing site must be able to monitor up to 24 ET Stations in any LCN.

The MEPS site implementation of CAT-ASVAB must also include a DHC unit. The main function of the DHC is to collect data daily from each LCN within the associated MEPS administrative segment, including LCN's located at MET sites. Data is transmitted to the DHC either over a hard-wired connection (in the case of MEPS LCN's) or modems in the case of MET LCN's.

MET Sites. At the MET sites, transportable computer systems will be used to administer CAT-ASVAB. The hardware configuration is to be based on the concept of a "generic" LCN. This generic LCN will consist of six ET Stations being monitored by a single TA Station, including any peripheral equipment. Note that many more (up to 24) ET Stations must be monitored by a single TA station and still maintain CAT-ASVAB performance requirements. In addition, it is important that the selected equipment support the CAT- ASVAB Stage 2 portability requirements; i.e., number of packages and weight requirements for a generic LCN (no more than eight components weighing a total of no more than 120 lbs., each component weighing 23 lbs). Environmental preformance requirements such as temperature, humidity, etc. must also be met.

The computer hardware configuration for an LCN may be described as follows: Each ET Station would consist of a response device, a screen display, and include access to sufficient RAM and/or data storage to permit the administration of any CAT-ASVAB test; the amount of RAM required depends on the particular application software and networking design being used to implement the functions. Each ET Station would be tied into a TA Station by networking cables; the TA Station being essentially an ET Station with a mass storage device and full-sized keyboard attached. Finally, a single(very portable) printer and modem for the TA Station would complete the complement of equipment composing the LCN.

The operational requirement for an LCN is to administer CAT-ASVAB to those military applicants scheduled for testing at the MET site. Initially, an Office of Personnel Management (OPM) examiner may be required to pick- up the LCN equipment at a staging area; if it were not secured at the testing site itself. The equipment would be transported to the test site, carried from the vehicle to the test site, and configured, ready for testing by the examiner. Once all examinees have completed testing, the examiner will attempt to telecommunicate examinee personal and test item response data to the DHC unit at the associated MEPS. This will be done using a modem and dial-up telephone line, if available at the test site. If this is not possible, the data will be transferred once the equipment is returned to the staging area.

Finally, if the equipment were not secured at the testing site, the examiner would package the equipment into transportable packages, carry these packages to a vehicle, and return the equipment to the staging area.

Design Considerations

It can be concluded from the survey of the requirements for developing a CAT-ASVAB System, that design efforts must be focused on the requirements for a generic LCN. This mainly includes the portability and functional capability of an LCN.

Portability. With respect to the portability aspect of an LCN, it should be clear that the response of the ACAP system to this requirement will be contingent on the capabilities of currently available commercial computer hardware. The ACAP System will not include the development and/or the building of computer systems to meet CAT-ASVAB needs, but rather will use commercially available computer hardware to support the accelerated field- test effort.

On-Line Data Storage. The on-line data storage requirement is the factor which will most influence the selection of computer hardware to support the ACAP System development. On-line data storage requirements will also significantly impact upon the design of the software for the ET, TA, and DHC units. The TA and ET Stations must have access to the on-line storage of two Forms of the CAT-ASVAB. Based on an estimate of 100 items per test, each Form will require approximately 850 Kilobytes (KB) of storage; assuming that the data is stored as a sequential file or equivalent. Actual data storage requirements would be 30-50% higher if the system design required that the data be stored as a random access file. Only one Form must be available to an ET Station during a test session. Additional storage for use by the TA and ET Station will have to be allocated for two experimental item pools, of 170 KB each (one pool for each form); an experimental item set derived from the experimental item pool, (10-125 KB); application software (approximately 250 KB per unit); two survey questionnaire item banks, 80 KB for both questionnaires; and examinee personal data such as information required on the USMEPCOM 714-A Form; and examinee personal and test item response data, 15 KB per examinee.

Test Administration Time. Depending on the network that is proposed to support ACAP, the time required to administer a CAT-ASVAB test could also affect the minimum response time required to have test data for the next item available at ET Stations.

Candidate Local CAT-ASVAB Network Designs

Using the preceding design considerations as a guide, the current commercial market in portable (or, at least, transportable) computer systems indicate that there are three basic designs upon which to develop a workable ACAP System. For the purpose and scope of the ACAP effort, the discussion will be focused upon the storage (and retrieval) capability of the candidates with respect to the 850 KB test item bank. It should be remembered that the important consideration is that the examinees have available (within the response time specified in the Stage 2 RFP), the correct test item as dictated by the underlying item selection strategy being used at the time.

The basic designs to be discussed are generic designs and, as such, may not be entirely represented by an actual example on the commercial market. The examples provided are those system designs that come very close to representing the generic design being discussed.

Design # 1 - CAT-ASVAB Items Stored on Removable Media

In this type of design, each ET Station would consist of sufficient internal storage on removable media (e.g., 3 1/2 in. micro-floppy diskettes) to accommodate the storage of the entire test item bank. Internal RAM would be about 512 KB. Necessarily, the test item bank files would have to be encrypted or "scrambled" on the removable media. The ET Station would be very portable and weigh from 11-17 lbs., and include a flat panel Liquid Crystal Display (LCD), electroluminescent, or equivalent low-weight display. The computer system that may currently best illustrate this hardware configuration is the Data General/One.

The main advantage to this type of configuration is that it is basically a very portable system. The total weight for a generic LCN could be as low as 80 lbs., including seven very transportable components (assuming that the LCN Networking requirement was suppressed).

There are many disadvantages to this design.

1. Security. The entire test item bank must reside on the removable media, necessarily jeopardizing the security of the test item bank files.

2. Media Updating. Each ET Station will require two removable diskettes installed in the disk drives to accommodate CAT-ASVAB testing. If this design were going to be installed at the ACAP field-test MEPS, and MET sites, approximately 400 ET Stations would be involved. This would require 800 micro-floppy diskettes to be inventoried and secured; a very large media creation, distribution, and security problem each time the test item bank is updated.

3. Ease of Use. Use of a removable storage media will require a significant amount of operator intervention to insert/remove diskettes. For each ET Station, two movements are required to "boot" the computer and to receive testing software. Eight diskette movements are required to transfer the examinee's response data to the examiner's work diskette; two movements can be avoided for subsequent examinees. A MET site LCN testing 10 individuals with six ET Stations available would require at least 100 movements.

4. Maintenance. Since the micro-floppy drives are in constant use during the testing process, system maintenance may be higher than some other configuration in terms of disk drive maintenance and diskette replacement. Each test item being displayed will require at least one disk access.

Design # 2 CAT-ASVAB Items Stored on a Central File Server

This type of LCN design is configured around a central file server (e.g., a hard disk) which acts as the repository for the CAT-ASVAB item banks and supporting data files. In this type of design, the capabilities of the network supporting the movement of data from the file server to each examinee station is of paramount importance. A minimal amount of RAM is available at the ET Station (less than 512 KB), with perhaps one internal floppy disk drive available. A central file server is required because each ET Station cannot support the entire requirement of data storage without adding significantly more components to the overall network. Typically, the ET Station is bulkier and may not support the latest flat screen technology; e.g., LCD or electroluminescent displays. The Macintosh Computer by Apple Corp. is an example of this type of design. In general, any configuration of equipment requiring a central file server would be an example of this design.

Perhaps the main advantage to this type of network is that a very sophisticated networking

capability must be installed in order to "make it work". Because this capability is available, one could, (theoretically) also install a TA Station monitoring capability. The monitoring capability would have to be installed in such a fashion so as not to compromise the response time requirements for the display of test items at the ET Stations in the LCN. Admittedly, the monitoring function could be potentially very useful at certain large MEPs, in which many examinees are being examined simultaneously. Unfortunately, this is also the situation in which the ET Station response time requirement would be most compromised. However, this system could be the least expensive of the networks being investigated. The cost of the file server could be distributed over each ET Station which could function without any removable media being available. Another advantage (as opposed to Design #1) is that the movement of test item bank data and/or examinee response data to/from the ET Station is automatic and does not require examiner intervention. Once the network is set-up and working properly, the examiner's tasks are minimized with respect to data movement requirements.

The main disadvantage to this design is concerned with the reliability of the file server itself; when the file server fails, the entire network is inoperable. To meet the CAT-ASVAB Stage 2 reliability requirements, it will be necessary to include two identical such in the LCN. This means that the system is heavy, environmentally intolerant, and requires a large number of components to be transported. In addition, for MET sites in which the equipment has to be assembled, and disassembled, during each testing session, a heavy requirement will be placed on the examiner to serve as a computer operator. This could conceivably result in a reclassification of the OPM examiner position.

Another disadvantage is that the system response time and monitoring requirements are functionally related in Design#2. It is very difficult to imagine a network operating system that can simultaneously accommodate these requirements in a fully loaded LCN; 24 ET Stations attached to a single TA Station. What is the maximum response time at any ET Station when all examinees are requesting items (simultaneously) from the same source (i.e., file server)? Note that sufficient time must also be allowed for de-encryption of the items before display, as well as decompression of graphics items, as necessary. In summary, another disadvantage of Design # 2 is that the maximum system response time is relatively large compared to other system designs, and therefore is a potentially compromising consideration relative to hardware selection for purposes of ACAP.

Design # 3 - Test Item Banks Stored in Examinee Test Station Random Access Memory

The TA and ET Stations in this design would consist of a large amount of internal RAM, on the order of at least 1.5 MB. The ET Station would be supported by one micro-floppy drive and probably include the latest in electroluminescent or LCD display. Therefore, for purposes of recovery due to network failure, the ET Station would be very responsive as it is capable of operating independently. In addition, as a networking capability will be available, the TA Station could perform the functions of an "electronic" file server. Total RAM available on the TA Station could be 1.5 to 3.5 MB (preferably higher); allowing for great flexibility in the total number of alternate forms available during any one test session.

Note that several removable media are required in order to "boot" the systems. However, a total of no more than TWO micro-floppy diskettes are required to store the test item banks (per Form) and supporting data files; each ET Station would also require one micro-floppy to be installed as a "working" diskette for failure recovery purposes. Normally (after initial "boot-up" at the beginning of a test session), no micro-floppy diskette movements are required by the examiner; i.e., the network would accomplish all data movements.

The main advantage of this design is that it offers a large degree of flexibility with respect to design options. The ET Stations are capable of operating as stand-alone devices and, as such, it is virtually impossible for an examinee's test session to fail to be completed; each ET

Station backs-up every other ET Station. For this reason, and as it minimizes accesses to a mechanical device, Design # 3 should be the most reliable of the network designs discussed. In addition, this design offers a very high level of security. Once power is removed from the computer, the volatile random access memory (RAM) is erased. This provides dependable security for the test item banks. Furthermore, as noted above, only two removable micro- floppy diskettes are required per form; regardless of the number of stations in the LCN.

Another important consideration, in comparison to Design # 2, is that the LCN monitoring and the system response time requirements are not functionally related. In addition, it is possible to configure a collection of computer hardware for Design # 3 that permits the entire 850 KB test item bank (for a form) to reside at the ET Station. Therefore, the response time required to display test items will be network independent. The item display process will be accomplished at RAM speed, resulting in a maximum response time that is on the order of milliseconds, as opposed to seconds.

Generally speaking, up to failure recovery, during actual test administration, the advantages of Design # 3 include: a) no removable media are requires, b) minimization of the need to use mechanical devices, c) a high level of security for test item banks, d) excellent system response time characteristics, and e) de-encryption of test item bank need occur but once; when this information is initially transferred to the "electronic" file server. In addition, some of the better features of Design # 2 are also characteristic of this design: a) examiner intervention to move data in an LCN is not required, and b) TA Station monitoring capability can be automated.

The primary disadvantage of this design is that it does tend to cost more than some alternatives. However, it is certainly true that the cost of computer hardware is decreasing. Another disadvantage is that there is only one viable candidate system on the market that would come close to exemplifying this design; that being the Hewlett-Packard Integral Computer. This system is currently capable of 1.5 MB of internal RAM (to include a networking interface card), with a single 3 1/2 in., 710 KB capacity, micro- floppy. Finally, in the case of the Integral, the ET Station would be somewhat heavier than some alternatives, weighing approximately 23 lbs. (assuming that the printer is removed).

Recommendations.

Of the three Designs discussed, the authors recommend Design # 3. They believe that it will give the Government the greatest amount of flexibility as the ACAP system is field-tested and the LCN configuration needs to be adjusted to accommodate future requirements. In addition, this design offers a system with the greatest amount of reliability and test item security. No mechanical devices are required to maintain normal CAT-ASVAB testing (up to recording examinee data on the micro-floppy associated with an ET Station for back-up purposes in the event that station fails during a test session). Test item bank data will be stored in volatile RAM instead of removable media, insuring the erasure of this sensitive information immediately when power is removed.

REFERENCES

(6 June 1984). Computerized Adaptive Testing Stage 2 Full Scale Development Request for Proposal issued by the Navy Regional Contracting Office, Long Beach.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hilldale, NJ:Erlbaum.

A Validity Study of the Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery

Kathleen Moreno & Daniel O. Segall
Navy Personnel Research & Development Center

William F. Kieckhaefer
RGI, Incorporated

PURPOSE

The purpose of this study was to assess the construct and predictive validity of a Computerized Adaptive Testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB). This study was conducted as part of an effort to evaluate CAT-ASVAB as a replacement for the paper-and-pencil ASVAB (P&P-ASVAB).

BACKGROUND

Over the past decade numerous empirical studies have been conducted to evaluate the construct and predictive validity of adaptive aptitude testing (Sympson and Moreno, 1985). Overall, the results of these studies indicate that adaptive testing is as valid as conventional, paper-and-pencil testing. However, the majority of these studies were conducted using verbal ability or arithmetic reasoning tests. Very little research has been conducted using aptitude tests measuring other types of ability.

The battery of interest in this study, the ASVAB, consists of tests measuring ten types of ability. The ASVAB is used by all military services for selection and classification of military applicants. This study examined the construct and predictive validity of all ASVAB tests.

APPROACH

Examinees

Examinees were military recruits scheduled for training in one of the military service specialties selected for inclusion in this study. Over all services, 7,515 examinees were tested. Sample sizes for each specialty are shown in Table 1.

Tests

P&P-ASVAB. The P&P-ASVAB is a group-administered, conventional battery in which all examinees answer all items in the same sequence. The P&P-ASVAB consists of eight power tests and two speeded tests: General Science (25),[1] Arithmetic Reasoning (30), Word Knowledge (35), Paragraph

[1] Values in parentheses are test lengths.

63

Comprehension (15), Auto and Shop Information (25), Mathematics Knowledge (25), Mechanical Comprehension (25), Electronics Information (20), Numerical Operations (50), and Coding Speed (84). There are six parallel forms of the P&P-ASVAB. This study used forms 8A and 9A. Number correct scores served as estimates of ability.

CAT-ASVAB. The CAT-ASVAB used in this study is an experimental version designed to measure the same abilities as those measured by the P&P-ASVAB. There are nine power tests and two speeded tests: General Science (197),[2] Arithmetic Reasoning (166), Word Knowledge (194), Paragraph Comprehension (95), Auto Information (179), Shop Information (189), Mathematics Knowledge (135), Mechanical Comprehension (70), Electronics Information (168), Numerical Operations (50), and Coding Speed (84). The nine power tests were administered adaptively using maximum information item selection and bayesian scoring. All power tests were terminated at a fixed length of 15 items, expect for Paragraph Comprehension, which was terminated after 10 items. The two speeded tests were administered in a conventional manner, with the test terminating after a fixed time. For the speeded tests, number correct scores were used to estimate ability.

Criterion Variables

Since each service has numerous training schools, prediction of performance for only a selected number could be assessed in this study. Schools were selected so that a wide variety of specialties would be represented. In addition, since military services use composites of test scores for selection and classification, schools were selected so that school composite scores would span all ASVAB tests. Table 1 lists the selected specialties and the criteria used for each specialty. For the majority of specialties, final school grade (FSG) or time to completion (TC) was used. However, for some specialties these measures were not available. In these cases, analyses were performed to determine which measure should be used.

Procedure

Examinees were tested approximately two weeks after arrival at a recruit training center, prior to entrance into training schools. Examinees were group-administered the experimental CAT-ASVAB and those P&P-ASVAB tests that were used in computing a recruit's school selection composite score. The tests were counter-balanced so that half the examinees took CAT-ASVAB first and half took P&P-ASVAB first. The CAT-ASVAB was administered using Sanyo monitors and Apple III computers networked with a Corvus hard-disk drive.

Pre-enlistment ASVAB scores were collected on all examinees from DD forms 1966. School performance data were collected after examinees had completed training.

---

[2] Values in parentheses are item pool sizes.

Data Analyses

Predictive Validity. For each of the selected specialties, composite scores were computed from CAT-ASVAB standardized scores and from P&P-ASVAB standardized scores. Validity coefficients were obtained for each test version by correlating school composite scores with school performance data. In order to test for significant difference between test versions, *t* values were computed. Since examinees in this study were a selected sample from the military applicant population, validity coefficients were corrected for range restriction using a multivariate approach (Lawley, 1943). No significance testing was performed using corrected validity coefficients.

Construct Validity. For each service, the intercorrelation matrix of CAT-ASVAB and P&P-ASVAB test scores was factor analyzed using the principal axes method, followed by a varimax rotation to simplify the factor structure.

RESULTS

Predictive Validity

Table 1 shows the validity coefficients obtained using CAT-ASVAB and P&P-ASVAB. Significance tests revealed no differences between the validity coefficients for the two test versions, even though CAT-ASVAB tests are much shorter than P&P-ASVAB tests.

Construct Validity

Table 2 shows the results of the factor analysis using data from the Air Force sample. Four factors were extracted, based on an eigenvalue of 1.0 or greater. These factors have been labeled as technical, verbal, mathematical, and speeded factors. As shown, the CAT-ASVAB tests had similar loadings to those of the corresponding P&P-ASVAB tests. Findings were similar for the other three services.

CONCLUSIONS

These results suggest that CAT-ASVAB is a viable alternate to P&P-ASVAB. In this study, CAT-ASVAB tests seem to be measuring the same abilities as the P&P-ASVAB tests and predict school performance as well as P&P-ASVAB tests, even though CAT-ASVAB test lengths are much shorter. However, before replacing the P&P-ASVAB with CAT-ASVAB, the two versions should be compared in terms of differential prediction by test version and subgroup membership. Such analyses are currently being performed at NPRDC.

Table 1

Validity Coefficients for CAT-ASVAB and P&P-ASVAB

Specialty	Criterion	Sample Size	Validity Coefficients CAT-ASVAB	P&P-ASVAB
NAVY				
Radioman	TC	186	-.41 (-.69)	-.40 (-.68)
Mess Management	FSG	170	.45 (.74)	.40 (.71)
Hospital Corpsman	FSG	192	.56 (.77)	.60 (.80)
Electronics Tech.	TC	143	-.41 (-.80)	-.46 (-.83)
Hull Maint. Tech.	FSG	170	.35 (.75)	.35 (.75)
Sonar Tech.	FSG	205	.46 (.79)	.46 (.80)
AIR FORCE				
Avionics	FSG	147	.56 (.80)	.56 (.80)
Administration	FSG	208	.22 (.57)	.15 (.53)
Aircraft Maint.	FSG	245	.54 (.81)	.49 (.79)
Medical	FSG	95	.64 (.87)	.62 (.86)
Security Police	FSG	456	.49 (.78)	.45 (.77)
MARINES				
Avionics	FSG	228	.49 (.81)	.48 (.81)
	TC	228	-.58 (-.84)	-.54 (-.84)
Administration				
Lejeune	FSG	72	.14 (.32)	-.05 (.13)
Pendelton	FSG	39	.22 (.49)	.29 (.56)
Aircraft Mech.				
School 1	FSG	181	.34 (.54)	.30 (.53)
School 2	FSG	69	.50 (.70)	.47 (.70)
Motor Transport	FSG	151	.29 (.47)	.30 (.48)
Combat Engineer	Sum(All)	123	.69 (.82)	.66 (.80)
Field Radio Opr.	Sum(1-4)	128	.33 (.43)	.21 (.37)
	Sum(5-8)	128	.06 (.46)	.09 (.47)
ARMY				
Infantry	Sum(All)	329	-.24 (-.34)	-.28 (-.36)
Mechanic				
Fort Dix	Average	198	.57 (.74)	.59 (.76)
Fort Jackson	PC	186	.35 (.52)	.38 (.54)
Motor Transport	Sum(All)	277	-.47 (-.63)	-.44 (-.63)
Administration	Wtd Sum	145	-.35 (-.64)	-.44 (-.72)
Telecom. Opr.	Sum(All)	169	.15 (.28)	.22 (.32)
Medical	FSG	225	.63 (.85)	.59 (.83)

Note. TC is the time for course completion; FSG is a final school grade; Sum(All) is a sum of training module scores; Sum(1-4) is a sum of scores on training modules one through four; Sum(5-8) is a sum of scores on training modules five through eight; Average is an average of the scores on all training school modules; PC is a percent correct score on the end-of-course test; and Wtd Sum is a sum of module scores minus a weighted typing score.

Numbers in parentheses are validity coefficients corrected for restriction in range.

Table 2

Results of a Factor Analysis of CAT-ASVAB
and P&P ASVAB Scores from an Air Force Sample

	Varimax Rotated Factor Matrix			
	Factor 1 (Tech)	Factor 2 (Verbal)	Factor 3 (Math)	Factor 4 (Speeded)
P&P-ASVAB				
AR	.28	.15	.66*	.31
WK	.17	.82*	.08	.05
PC	.16	.50*	.13	.16
NO	-.07	.00	.24	.70*
GS	.40*	.63*	.22	-.01
CS	.00	.01	.06	.71*
AS	.82*	.14	-.01	.00
MK	.17	.25	.80*	.23
MC	.65*	.21	.36*	-.01
EI	.61*	.32*	.19	-.03
CAT-ASVAB				
AR	.31*	.27	.71*	.22
WK	.15	.85*	.16	.03
PC	.17	.68*	.23	.13
NO	-.03	.13	.26	.65*
GS	.33*	.73*	.28	.00
CS	-.03	.10	.08	.71*
AS	.90*	.15	.09	-.02
MK	.08	.29	.74*	.23
MC	.66*	.22	.30*	-.03
EI	.64*	.42*	.26	-.08

*factor loading > .30

REFERENCES

Lawley, D. (1943). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh Proceedings, Section A, 62*, 28-30.

Sympson, J. B., & Moreno, K. E. (1985, August). *Validity of adaptive testing: A summary of research results*. Paper presented at the American Psychological Association Convention, Los Angeles, CA.

Assessment of the Unidimensionality of CAT-ASVAB Subtests

Mary K. Schratz
Navy Personnel Research and Development Center
San Diego, California 92152-6800

Background

A joint-service project is underway to develop a computerized adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). CAT-ASVAB is intended to replace the paper-and-pencil ASVAB (P&P-ASVAB), used by the four military services to select and classify applicants. The present P&P-ASVAB battery consists of ten aptitude subtests. These subtests, and the number of items included in each for a given form of P&P-ASVAB, are listed in Table 1 below.

Table 1. P&P-ASVAB Subtests and the Number of Items Included per Form

Subtest	Number of Items
General Science (GS)	25
Arithmetic Reasoning (AR)	30
Word Knowledge (WK)	35
Paragraph Comprehension (PC)	15
Numerical Operations (NO)	50
Coding Speed (CS)	84
Auto and Shop Information (AS)	25
Mathematics Knowledge (MK)	25
Mechanical Comprehension (MC)	25
Electronics Information (EI)	20

The Numerical Operations and Coding Speed subtests are speeded measures; all other subtests are power measures. All examinees taking a given P&P-ASVAB form are administered the same set of test items in all subtest areas.

While CAT-ASVAB is intended to measure the same subtest areas as P&P-ASVAB, the power subtests of CAT-ASVAB will be administered adaptively. Within each power subtest, different items will be selected for computer administration to examinees, depending on their performance on previously administered items. The testing process is thus individualized for examinees.

The Problem

The theoretical framework supporting the adaptive testing process to be implemented in CAT-ASVAB is item response theory (Lord, 1980). The item response theory methods to be used assume that each subtest is unidimensional. The outcome of this assumption is that each item included in a given subtest should measure the same unitary construct, in addition to having specific and error variance components associated with it. The Committee for an Evaluation Plan for the Computerized Adaptive Vocational Aptitude Battery (Green, Bock, Humphreys, Linn, & Reckase, 1982) states that unidimensionality is always advisable for tests of ability, but is more important for adaptive tests. Though some may be of the opinion that the IRT model to be applied in CAT-ASVAB is strong enough to counteract potential problems with respect

to multidimensionality in CAT-ASVAB subtests, further study of this problem is necessary. Unidimensionality may be considered an important problem for an adaptive test because different items are administered to examinees. For a traditional test such as the P&P-ASVAB, all examinees are presented the same test items, and are given an equal opportunity to respond to them irrespective of their dimensionality. The individualized nature of adaptive tests may raise questions related to test fairness. The present paper describes the approach to be taken, progress, and plans for dimensionality analyses of CAT-ASVAB items.

Approach

A recent advance in the development of factor-analytic approaches to exploring test dimensionality is "full-information item factor analysis." Bock, Gibbons, and Muraki (1985) contend that of the various methods that have been proposed for investigating dimensionality of item sets, item factor analysis is the most sensitive and informative. This method of item factor analysis is based on item response theory; it uses all data as distinct item response vectors. Thurstone's multiple factor model is used. The procedure is implemented by marginal maximum likelihood estimation and the EM algorithm. Statistical significance of the addition of successive factors to the model is tested by a likelihood ratio criterion. Provisions for the effects of guessing on multiple choice items, and for omitted and not reached items, are included.

One of the applications of this methodology to real data, presented by the authors as evidence for the accuracy and practical utility of the method, is an analysis of the power subtests of P&P-ASVAB. The analysis was conducted in a ten-percent random sample of data from the Profile of American Youth Study. The number of cases used in the analysis was 1,178, drawn from a total sample of 11,817 subjects. The details of the item factor analyses are presented in the Bock et al. (1985) report. For the purposes of this presentation, the results obtained for P&P-ASVAB are relevant to the selection of and application of an approach for studying test dimensionality in the intended replacement battery, CAT-ASVAB. Thus they will be described briefly.

For the P&P-ASVAB General Science test, Bock et al. (1985) found two significant factors; one factor was interpreted as a physical science factor and the other factor was interpreted as a biological (or health science) factor. Two factors were also found for the Arithmetic Reasoning subtest. While the second factor found was a minor one, the authors have interpreted it as a business arithmetic factor. For the Word Knowledge subtest, clear evidence for a second factor was found, though the factor has no apparent relationship to item content. For the Auto and Shop Information subtest, the authors found clear evidence for two factors separating the two types of items. For the Mathematics Knowledge subtest, two significant factors were also found; one factor involved items requiring knowledge of formal algebra and the second factor involved numerical calculation and mathematical reasoning. Only one factor was found for the Paragraph Comprehension, Mechanical Comprehension, and Electronics Information subtests.

Bock et al. (1985) conclude that the applications of the procedure reported in their paper show that, for moderately large samples, minor factors can be detected. The procedure is recommended as an exploratory technique in searching for item features that are responsible for individual differences in cognitive test performance.

Given the reported adequacy and practical utility of the full-information factor analysis approach to the detection of multidimensionality in item sets, and the indications of multiple factors found by Bock et al. (1985) in the subtests of the P&P-ASVAB battery which CAT-ASVAB is intended to replace, NPRDC plans to make use of this procedure in studying the dimensionality of the CAT-ASVAB items.

Progress and Plans

One of the requirements of the adaptive testing process is the development and calibration of a large bank of test questions, covering a wide range of difficulty for the intended test-taking population, for use in the item selection process. In a study carried out by the Air Force Human Resources Laboratory, and conducted under contract by Assessment Systems Corporation, a bank of 2,118 test items intended for operational use in CAT-ASVAB was developed and calibrated in 1983. This bank of test questions covers all of the P&P-ASVAB power subtest areas, with more than 200 items developed for each subtest. On the basis of statistical and judgemental criteria, items will be selected from the total number available for inclusion in the final CAT-ASVAB battery. One of the criteria for accepting an item for inclusion in the final battery is measurement of a one-dimensional universe represented by a pool of items. This pool may represent a subtest, or, alternatively, a subset of items within a subtest. From unidimensional pools of items, individual test items will, of course, be selected for administration to examinees in the adaptive testing process.

Preliminary full-information item factor analyses of one of the more suspect subtest areas of the CAT-ASVAB item bank, given the results of Bock et al. (1985), have been conducted. Analyses of the General Science items indicate that indeed the item factor analysis procedure is sensitive to the presence of minor factors in the data. The CAT-ASVAB General Science items were developed to measure three main subject areas: (1) Life Science; (2) Physical Science; and (3) Earth Science (Prestwood, Vale, Massey, & Welsh, 1984). Items were randomly assigned to four test booklets for calibration purposes and approximately 2,500 examinees were tested with each booklet. Inspection of the individual items, the content categories which they were written to represent, and the preliminary item factor analysis results suggests some clustering of items in terms of subject matter. Where a preponderance of items of one type have been assigned to a booklet, the procedure appears to be sensitive to the detection of minor factors.

It is NPRDC's intention to conduct full-information item factor analyses for each power subtest to be included in the CAT-ASVAB battery. This will be done in a joint calibration analysis of both the CAT-ASVAB items and P&P-ASVAB items for each subtest. While CAT-ASVAB is intended to replace P&P-ASVAB, both batteries will concurrently be administered in an operational setting. CAT-ASVAB subtest scores must be scaled to those of P&P-ASVAB and, as is the case for the present P&P-ASVAB, a single CAT-ASVAB score will be generated for each subtest. The joint item factor analysis of CAT-ASVAB items and P&P-ASVAB items for each subtest is expected to result in as many or more factors than those determined for the corresponding P&P-ASVAB subtest alone. Where more than one dimension is present, the joint calibration will allow for transformation of the resulting item parameters to congruence with those parameter estimates obtained from the analysis of P&P-ASVAB items alone. Dr. Bruce Bloxom of Vanderbilt University is presently working on a procedure for combining *m* multidimensional ability

scores into a single score comparable to that obtained on P&P-ASVAB.

Recommendations

The individualized nature of adaptive tests raises interesting test development issues related to test fairness. The adaptive testing process to be implemented in CAT-ASVAB is supported by item response theory methods which assume that a single underlying trait is measured within a subtest. Bock et al. (1985) have provided a promising procedure for investigating conformance to this assumption. It is recommended that the full-information item factor analysis approach, currently being used as an exploratory technique in the development of CAT-ASVAB item pools, be considered for use in other test development applications involving item response theory methods.

REFERENCES

Bock, R.D., Gibbons, R., & Muracki, E. (1985, August). Full-information item factor analysis (MRC Report 85-1). Chicago, ILL: Methodology Research Center/NORC.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.B., & Reckase, M.D. (1982, May). Evaluation plan for the Computerized Adaptive Vocational Aptitude Battery (Research Report 82-1). Baltimore, MD: The Johns Hopkins University, Department of Psychology.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Prestwood, J.S., Vale, C.D., Massey, R.H., & Welsh, J.R. (1985, September). Development of an adaptive item pool for the ASVAB (AFHRL-TR-85-19). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

SEX DIFFERENCES IN IRT TRUE-SCORE EQUATING

Daniel O. Segall
Navy Personnel Research and Development Center

William F. Kieckhaefer
RGI, Incorporated

Kathleen E. Moreno
Navy Personnel Research and Development Center

Introduction

The current Armed Services Vocational Aptitude Battery (P&P-ASVAB) is a paper-and-pencil test with a fixed sequence of test items. The Navy Personnel Research and Development Center is developing a computerized adaptive version (CAT-ASVAB) as a possible replacement for that test. The CAT-ASVAB will tailor the difficulty of the test items administered to the individual from responses to earlier items. This testing method is expected to increase the efficiency of selecting and classifying new accessions.

If the CAT-ASVAB becomes operational, it will be implemented gradually so that some examinees will be administered the CAT-ASVAB while others will receive the P&P-ASVAB. The two versions use different estimators of ability. The P&P-ASVAB uses a number correct score for each examinee, while the CAT-ASVAB computes a bayesian estimate of ability. Scores from the two versions must be equated so that personnel selection and classification decisions do not vary between test versions.

Braun and Holland (1982) give one definition of test equating. They adopt the definition Form-X and Form-Y are equated on population P if the distribution of the transformed y scores in population P is the same as the distribution of the untransformed x scores. Applying this definition to the current problem, the CAT-ASVAB and P&P-ASVAB are equated on population P if the distribution of the transformed CAT-ASVAB scores in this population is the same as the distribution of the P&P-ASVAB scores. This definition has the desirable quality of assuring equal flow rates for the two versions of the ASVAB.

Unfortunately two tests that are equated on population P may not be equated for various subpopulations that are included in P. Test scores that are equated for the military applicant population may not be equated for either the population of female applicants, or the population of male applicants.

This paper investigates the application of IRT true-score equating to the experimental CAT-ASVAB. An effort is made to determine whether CAT-ASVAB scores can be transformed to a paper-and-pencil scale without placing either males or females at a disadvantage relative to their P&P-ASVAB scores.

Method

SUBJECTS. During April of 1984, 200 male and 200 female Army recruits at Fort Jackson, South Carolina participated in this study. Each subject took both the CAT-ASVAB and the P&P-ASVAB.

SUBTESTS. The five subtests of the P&P-ASVAB selected for this study were taken from Form 8a: Arithmetic Reasoning (AR), Word Knowledge (WK), General Science (GS), Paragraph Comprehension (PC), and Numerical Operations (NO).

There were five CAT-ASVAB subtests designed to measure the same aptitude as the P&P-ASVAB subtests mentioned above: AR, WK, PC, NO, and GS.

PROCEDURES. Proctors seated all subjects in the testing area for the ASVAB and instructed them to complete the Privacy Act Statement. Then, half the subjects took seats in the adjacent CAT-ASVAB testing area. Subjects in this condition completed the CAT-ASVAB first and the ASVAB second. Subjects in the other condition (i.e., the remaining half of the subjects) took the P&P-ASVAB first and the CAT-ASVAB second.

Equating Transformations

Although each CAT-ASVAB subtest is designed to measure the same cognitive ability as its P&P-ASVAB counterpart, the two versions are not on comparable scales. The CAT-ASVAB power subtests produce an ability estimate ($\hat{\theta}$) while the P&P-ASVAB produces a number correct score (x). Although the CAT-ASVAB subtest of Numerical Operations does produce a number-correct score similar to the P&P-ASVAB, the method of responding has been shown to effect the CAT-NO score distribution. Thus for each content area, some method of equating the two versions is necessary.

We used two different procedures to equate the five subtests, depending on whether the subtest was adaptive or speeded. We used a procedure similar to one recommended by Green, Bock, Linn, and Recakase (1985) to equate the CAT-ASVAB power subtests. This procedure transforms the thetas into expanded expected number correct (EENC) scores. We used an equipercentile method to equate the CAT-NO speeded subtest. Both procedures are described in detail below.

POWER-SUBTESTS. For the power-subtests scores, we performed a two-stage equating. First we calculated expected number correct scores for the four CAT-ASVAB power subtests. Equation (1) transformed the estimated CAT ability, $\hat{\theta}_a$, for each person a, to an expected number correct score $\hat{\xi}_a$.

$$\hat{\xi}_a = \frac{1}{6} \sum_{k=1}^{6} \sum_{i=1}^{n} P(a_{ik}, b_{ik}, c_{ik}; \hat{\theta}_a), \qquad (1)$$

where n equals the number of items in the P&P subtest, and $P(a_{ik}, b_{ik}, c_{ik}; \hat{\theta}_a)$ represents the item characteristic curve defined by the three parameter logistic model evaluated at $\hat{\theta}_a$, for item i of P&P-ASVAB form k (where $k = 1,2,...,5,6$)

We substituted item parameter estimates $\hat{a}_{ik}$, $\hat{b}_{ik}$, $\hat{c}_{ik}$ (Sympson & Hartmann, 1985) for the values a_{ik}, b_{ik}, c_{ik} in equation (1). Then for each estimated ability, $\hat{\theta}_a$, an expected number correct score, $\hat{\xi}_a$, was obtained from (1).

Second, we applied a linear transformation to the scores computed from equation (1). This produced expanded expected number correct (EENC) scores that possessed the same mean and variance as the observed scores of the corresponding P&P-ASVAB. Finally, we rounded these EENC scores to the nearest integer value to produce the CAT-EENC scores. We repeated all the above procedures for each of the CAT-ASVAB power subtests (AR, WK, PC, and GS).

SPEEDED-SUBTEST. Numerical Operations (NO) was the only speeded subtest included in this study. This subtest is not adaptive and produces a number correct

score.

We obtained NO-equating data from 1,364 Army recruits. Each recruit received both versions of the subtest: (a) the CAT-NO subtest, and (b) forms 8 or 9 of the P&P-NO subtest.

We used an equipercentile method to equate CAT-NO to P&P-NO. First, we obtained CAT-NO and P&P-NO number correct scores at 99 different cumulative-percentile points. These cumulative-percentile points were obtained at unit intervals ranging from 1 to 99, inclusive. Next, we used least-squares-polynomial regression to smooth the equipercentile-equating function. We calculated polynomial regressions of several different orders and judged the quintic regression to provide the best fit based on a root-mean-squared-error criterion. We equated scores below the second percentile (of the CAT-NO score distribution) using linear interpolation from the (0,0) point to the point corresponding to the second percentile. Then we obtained smoothed-equipercentile estimates for each number-correct score using the estimated-polynomial-regression equation (or by linear extrapolation). Finally, these values were rounded to the nearest integer.

Results

Kolmogorov-Smirnov two-sample (KS) tests were used to test the difference between the P&P-ASVAB and equated CAT-ASVAB distribution functions. KS tests were run for all five subtests.

The total sample was first randomly divided into two groups, with the restriction that total group size was approximately equal and the number of females and males did not differ by more than one across the two groups. The next step computed expected number correct (ENC) scores from CAT-ASVAB thetas for each group. The linear transformation which transforms ENC scores to EENC scores was estimated for each group separately. Each transformation was then used to compute the CAT-EENC scores for that group.

Two comparisons were made: (1) a comparison of the CAT-EENC score distribution of Group A to the P&P-ASVAB score distributions of Group B and, (2) a comparison of the CAT-EENC score distribution of Group B to the P&P-ASVAB score distributions of Group A. Differences were tested using the KS statistic. The above procedure was repeated separately for the male, female, and combined samples.

Table 1 presents the results of the KS tests. Each comparison examines the difference between the CAT-EENC (or number correct for the NO subtest) distribution function and the corresponding P&P-ASVAB distribution function. Only two comparisons were significant at the .05 level: (1) the comparison of scores on NO for females, and (2) the comparison involving the AFQT composite for females.

Discussion

The results of the KS analysis indicate that the IRT true-score equating procedure provides similar distributions for the two versions of the ASVAB. Neither males or females appear to be placed at a disadvantage relative to their P&P-ASVAB scores.

References

Braun, H. I., and Holland, P. W. (1982). Observed-score test equating: a mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin

(Ed.), *Test Equating*. New York: Academic Press, 9-50.

Green, B. F., Bock, R. D., Linn, R. L., Lord, F. M., & Reckase, M. D. (1985). *A plan for scaling the Computerized Adaptive Armed Services Vocational Aptitude Battery (ASVAB)* (MPL TN 85-2). San Diego, CA: Manpower and Personnel Laboratory, Navy Personnel Research and Development Center.

Sympson, J. B. & Hartmann, L. (April, 1985). Item calibrations for Computerized Adaptive Testing (CAT) experimental item pools. In D. J. Weiss (Ed). *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Table 1
Kolmogorov-Smirnov Two-Sample Tests

Sex	Subtest	Comparison One		Comparison Two	
		Z	Probability	Z	Probability
Males	AR	.927	.36	.480	.98
	WK	.758	.61	.478	.98
n_1=72	PC	.677	.75	.541	.93
n_2=73	NO	.883	.42	.500	.96
	GS	.816	.52	.715	.69
	AFQT	.926	.36	.500	.96
Females	AR	1.178	.13	.478	.98
	WK	.970	.30	.904	.39
n_1=99	PC	.460	.98	.547	.93
n_2=100	NO	1.679	.01	.676	.75
	GS	1.078	.20	.901	.39
	AFQT	1.444	.03	.600	.86
Combined	AR	.737	.65	.517	.95
	WK	.884	.42	.948	.33
n_1=171	PC	.472	.98	.642	.81
n_2=173	NO	1.223	.10	.684	.74
	GS	.476	.98	.845	.47
	AFQT	.802	.54	.474	.98

Reducing the Predictability of Adaptive Item Sequences

C. Douglas Wetzel

Navy Personnel
Research & Development Center
San Diego, CA 92152

James R. McBride

The Psychological Corporation
San Diego, CA 92101

Previous research into the psychometric properties of Computerized Adaptive Testing (CAT) has shown that adaptive tests are much more efficient than conventional tests (e.g., Weiss, 1974, 1982). Different methods of adaptive testing vary widely in efficiency. The most efficient are those in which test items are chosen one at a time, in a manner which optimizes some function of the difference between item difficulty and the current estimate of the examinee's ability. However, many optimization CAT strategies yield a predictable sequence of test items early in testing that could lead to over exposure of items and possible compromise. This is because the possible sequences of test items form a binary tree. The same item will always be chosen first; only two items can be chosen second, and so on. As a result, all examinees who answer the first several items the same way will encounter identical sequences of test items, making compromise easy and almost inevitable.

This repetition of predictable item sequences is illustrated in Figure 1 with plots of individual examinee ability estimates as a function of test length. The course of testing for each examinee is traced up the page, with all examinees starting at an ability estimate of zero prior to being given their first test question (i.e. item zero). It can be seen that after the first test item is administered, there are only two possible ability estimates from right or wrong responses, and after the second item there are just four possible ability estimates, and so forth. As the number of items increases, the common paths shown early in testing fan out into a number of unique ability estimates. Smaller changes in the ability estimates are found later in testing (items 10-15) where they "home in" on a region of the ability continuum and become more reliable.

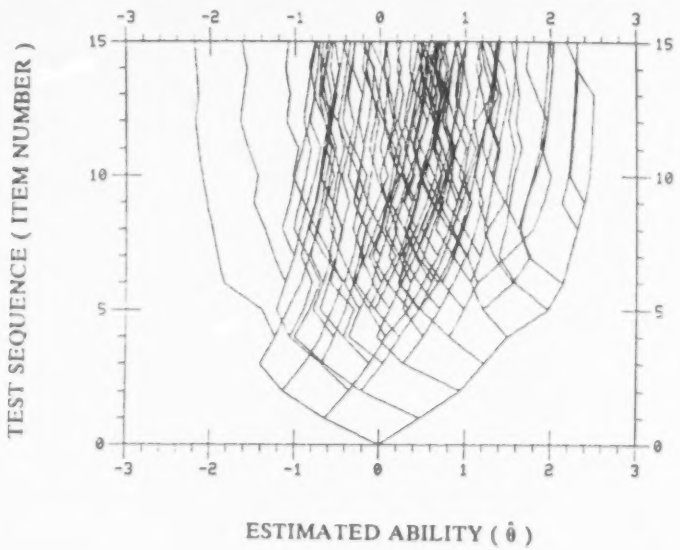


Figure 1. Item-by-item ability estimates ($\hat{\theta}$) for 75 Marine Corps recruits given a 15 item adaptive test using Owen's Bayesian strategy (Owen, 1969, 1975).

This situation suggests the need for a method which retains the efficiency of the mathematically optimal adaptive strategies, but one which eliminates the occurrence of predictable sequences of test items. One method of avoiding predictable item sequences is to choose an item at random from among a set of nearly optimal items. In the present work, a stratified maximum information (STMI) strategy was selected for extensive study of random selection from several good or informative items taken at different points in the sequence of items. A question to address is whether this technique to reduce the repeated exposure of the very best items in the bank would reduce the psychometric quality of the resultant adaptive test, relative to several other adaptive and conventional test strategies.

APPROACH

A two-stage computer simulation was used to investigate the effect of various item selection strategies on the psychometric characteristics of the resultant tests. The first stage simulation generated item parameter estimates with typical error characteristics. This item bank was used for the second stage in simulated administrations of adaptive and conventional tests for normal and rectangular examinee true ability distributions.

Generating Fallible Item Parameter Estimates: A simulated item bank was created based on real item parameters representative of a "live" testing situation. The item bank consisted of two parameters sets: the "true" parameters $\{ a , b , c \}$, and simulated estimates $\{ \hat{a} , \hat{b} , \hat{c} \}$. Unlike many synthetic item banks (e.g. Wetzel & McBride, 1983), real items written for live examinees often yield unique item banks without uniform distributions when broken down by each item parameter, and may have a distinct positive correlation between the $\hat{a}$-parameter and $\hat{b}$-parameter $r_{\hat{a}\hat{b}}$ (Sympson, Weiss and Ree; 1982). To achieve a test information curve that approximated that of real test items, a simulated bank was based on item parameters from real test items. Estimated item parameters from real items were used as true parameters for a simulation of the item calibration phase. The 200 real item parameters were obtained from J.B. Sympson of NPRDC from his calibration of five ASVAB content areas: word knowledge, arithmetic reasoning, paragraph comprehension, general science and, mathematical knowledge. These banks had each been calibrated with LOGIST (Wood, *et. al.*, 1976) on approximately 1500-2000 live examinees for entrance into the armed forces. In the present study, a stratified random sample from these five banks (970 total items, from individual banks of 180-210) was taken by randomly selecting 40 items from each of the five banks to yield a new combined total of 200 items. These 3-parameter logistic estimates were then used *as if* they were 'true' item parameters in a simulation in which examinees were administered all 200 items. Examinee item responses were simulated by using the 3-parameter logistic model (Birnbaum, 1968) to generate simulated binary responses to the test items using a probability sampling technique often employed for this purpose (Vale and Weiss, 1975). If a random number drawn from a uniform distribution on the interval (0,1) was less than the 3-parameter logistic model probability of a correct response $P(\theta)$, then the examinee was credited with a correct answer; otherwise an incorrect item response was specified. The simulated correct and incorrect item responses were created for 1500 normally distributed simulated examinees for these 200 items and then calibrated for the present study with LOGIST. These parameter estimates were then joined with the generating "true" parameters for simulations of the various test strategies studied here.

Examinee True Ability (θ) Distributions: Each test strategy simulation run was conducted twice with 1900 simulated examinees: (1) once with a *Rectangular θ Distribution* of 19 groups (100 examinees each) .25 θ units apart over the the interval

-2.25 to +2.25 ($\bar{X}$ = 0.0, S.D. = 1.369), and (2) once with one group of 1900 from a standard *Normal θ Distribution* ($\bar{X}$ = 0.0001, S.D. = 0.999).

Simulation of Testing Strategies: A fixed test length of 15 items was selected as a representative number of items in which an adaptive test could possibly be implemented. For each test strategy, 1900 examinees were simulated from the two ability distributions. Examinees always responded to the true item parameters { a , b , c } on the basis of true ability (θ) according the 3-parameter logistic model. An item-response was scored correct by the probability sampling technique described above. Item selection was based on the estimated parameters { $\hat{a}$, $\hat{b}$, $\hat{c}$ } . Ability estimation in the two Bayesian tests (see below) used the same estimated item parameters as used for item selection. All adaptive tests began with an initial ability estimate ($\hat{\theta}$) of 0.0 and the two Bayesian tests assumed an initial normal prior distribution of ability with variance 1.0. Either Bayesian or maximum likelihood ability estimates were limited to the range + 3.00 to - 3.00. Each testing strategy is described below.

Owen's Sequential Bayesian Test: This adaptive strategy chooses an item to minimize the expected value of the posterior variance of the ability distribution (Owen, 1969, 1975). The ability estimate (distribution) is updated after each item and the parameters of the Bayes posterior distribution are used as parameters of the prior distribution for the next item.

Stratified Maximum Information Test - Bayesian Scoring (STMI-B): This adaptive test selects the item with approximately the greatest item information ($I(\hat{\theta})$) at the current Bayesian ability estimate $\hat{\theta}$. Items are selected from a prearranged "information table" consisting of stratified lists of information values calculated for fixed θ levels. Each list of the information table contains information values arranged in descending order of the values of their information functions at the midpoints of a series of narrow intervals of ability. In this study, 36 lists in .125 wide θ increments spanned the ability range from -2.25 to +2.25. The ability estimate was updated after each item with the same Bayesian ability estimation procedure employed in the Owen's test above. This STMI-B strategy is a 'hybrid' (Wetzel & McBride, 1983) between previous strategies in that the same information table method is employed, but Bayesian ability scoring is used instead of maximum likelihood scoring (Sympson, Weiss & Ree, 1982).

Seven versions of the STMI-B test were simulated to investigate probabilistic item sequences produced by randomizing the choice among items. The top k consecutive items with greatest information in an information table's list were selected and held in a temporary vector. One of these k items was then selected randomly, with each of the items having equal $1/k$ selection probabilities. This randomization among the most informative items within a given list of an information table occurred in the list closest to the current ability estimate. Two types of randomization conditions were studied, which differed in whether the $1/k$ probability was constant for all 15 test items or varied as a function of the order of item administration:

Constant Selection Ratios for All Items Administered:

(a) 1/1 1/1 (b) 1/5 1/5 (c) 1/10 1/10 (d) 1/20 1/20 (e) 1/40 1/40

Selection Ratios Adjusted According to Test Length:

(f) 1/5 1/4 1/3 1/2 1/1 1/1 (g) 1/10 1/8 1/6 1/4 1/2 1/2

The constant selection ratios remained the same for each of 15 items administered (a-e). The first adjusting selection ratio strategy (f) selected the first item in the test from the best five available items in the current information table list (1/5), the second item from four (1/4), the third from three (1/3), the fourth from only two (1/2), and

then the fifth through the fifteenth items were each the single most informative item available (1/1). The second adjusting ratio strategy (g) used the first four ratios shown for the first four items and then used a ratio of 1/2 for the fifth through the fifteenth items. These adjusting ratio strategies randomize more for early items since they are most subject to compromise. Randomization decreases as the test proceeds so the strategy will select more appropriate items when the ability estimate is closest to it's terminal value.

Maximum Information Full Search - Bayesian Scoring (MI-B): Each item is selected by a search of the entire bank for the item with the greatest item information at the exact value of the current estimated ability ($\hat{\theta}$). This exhaustive item-by-item search is made by actually calculating item information for $\hat{\theta}$ "on-line" each time an item is to be selected during test administration. This strategy was created to assess any effect of granularity in the STMI-B test, where the θ continuum was divided into discrete increments in rounding $\hat{\theta}$ to the nearest .125 midpoint. Previous studies using the MI strategy have employed maximum likelihood scoring rather than the present Bayesian scoring (e.g., McKinley & Reckase, 1981; Weiss, 1982).

Stratified Maximum Information Test - Maximum Likelihood Scoring (STMI-ML): This strategy employs Bayesian ability estimation until at least one correct and one incorrect item response have been obtained and, then uses maximum likelihood estimation for the remainder of the items in the test. Items were selected from an information table calculated over the same .125 wide θ increments used for the STMI-B strategy above. This strategy was first used by Sympson, et. al. (1982).

Weiss's Stradaptive Test: This mechanical adaptive strategy uses a pre-sorted item bank divided into 'strata' on the basis of the $\hat{b}$-parameter and then arranges items within each stratum in descending order of the values of their $\hat{a}$-parameters (cf. Weiss, 1974). There were nine strata in this study, each 0.5 ability units wide, over the range -2.25 to +2.25. Items were selected from the top of the stack in each stratum. Branching to another stratum occurred after each item; branching up one stratum after a correct response, and down one stratum after an incorrect response.

Peaked Conventional Test: This conventional test was designed by selecting the 15 items with the greatest values of information at the central ability value of 0.0. All simulated examinees were administered this same set of 15 items.

Flat Conventional Test: This test was created by selecting the item with the greatest item information value at each of 15 equally distributed ability points over the interval -2.0 to +2.0. All simulated examinees were administered this same set of 15 items, with item difficulty increasing as the test proceeded.

RESULTS

Fidelity: The correlation or fidelity between true and estimated ability ($r_{\theta\hat{\theta}}$) for the simulated test strategies was based on the normal θ distribution consisting of a single group of 1900 examinees. All coefficients were based on Bayesian ability estimates, except the STMI-ML strategy which used maximum likelihood ability estimation. All the optimization strategies yield similar fidelities (Owen's Bayesian {.951}, STMI-ML {.950}, STMI-B {.953}, MI-B {.956}), with the Stradaptive test {.935} performing slightly better than the conventional tests. The peaked conventional test {.893} is lower than the flat conventional {.922} , since the peaked test does not span the extremes of ability. As the denominator of the fixed selection ratio increases, the STMI-B test shows a small decrease in fidelity {.953, .955, .952, .951, .936}, but at 1/40 it never falls below the conventional tests. The adjusting ratio STMI-B strategies yield fidelities { both .957

} comparable to the 1/1 and 1/5 fixed ratios and to the MI-B full search strategy.

Average Test Information: Test information was employed as an index of precision, or of how well a set of items discriminates an ability level from nearby ability levels; the reciprocal of the square root of information is inversely related to the standard error of an ability estimate (Lord, 1980). It was used here as a measure of the appropriateness of the set of items administered for a given true ability level. "Test information" $\{\sum I(\theta)\}$ is the sum of the individual "item information" $\{I(\theta)\}$ values for the items administered in an individual examinee's test. The true item parameters and true examinee ability were used to compute these test information values, which were then averaged over the 100 examinees at each of the 19 levels of true ability. Figure 2 shows the obtained test information (Lord, 1980) averaged over the 100 examinees in each of the 19 rectangular θ distribution groups for each test. A peak is shown around a $\theta \approx +1.25$ which is owed to the correlation of the $\hat{a}$ - and $\hat{b}$ -parameters ($r_{\hat{a}\hat{b}} = .577$, $r_{\hat{a}\hat{c}} = .059$, $r_{\hat{b}\hat{c}} = -.134$). The magnitude of the $r_{\hat{a}\hat{b}}$ correlation was very similar to that obtained by Sympson, et. al. (1982).

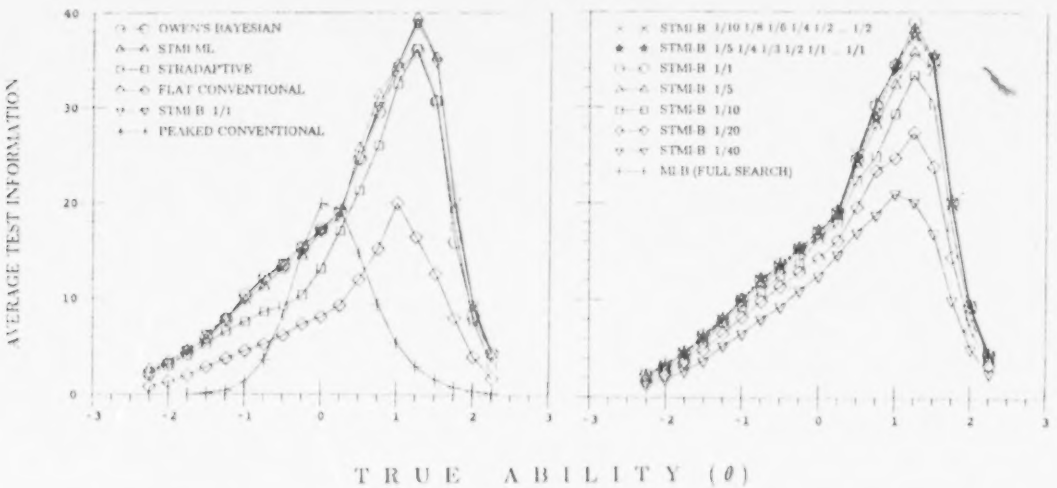


Figure 2. Average test information for test strategies over 19 true ability (θ) groups.

The left-hand panel of Figure 7 shows the adaptive tests to yield more test information than the two conventional tests over a wide range of θ , excepting the peaked conventional test at the narrow region around $\theta \approx 0$. The conventional tests represent one extreme of mismatch, where the same fixed set of items are presented regardless of the location of the examinee on the ability continuum. The Stradaptive test generally yielded less test information than the other adaptive tests at lower ability levels because it used a mechanical strategy that did not correct for guessing. The 1/1 STMI-B and STMI-ML tests yielded the best test information overall and were practically equivalent to the Owen's Bayesian test.

The right-hand panel of Figure 2 shows eight versions of the maximum information strategy using Bayesian scoring. For the five constant ratio STMI-B strategies { 1/1, 1/5, 1/10, 1/20 & 1/40 }, test information is degraded monotonically as the denominator of the ratio increases, i.e., the selection set included items farther down in the information table which had somewhat lower discriminations, higher guessing and more inappropriate difficulties. As the ratio changes from 1/1 to 1/40, the amount of

obtained change becomes larger for each doubling of the denominator. Substantial reductions were produced with the extreme 1/20 and 1/40 ratios, and a small but acceptable degradation in test precision resulted when 1/5 available items was randomly selected throughout the test.

The remaining tests shown in the right-hand figure panel all yielded test information that was approximately the same as the STMI-B 1/1 condition and in excess of the STMI-B 1/5 test. First, the MI-B condition achieved no more test information than any other condition, indicating that the .125 wide increments on which the STMI information table tests were based were small enough to closely approximate this full search condition. Second, the two STMI-B adjusting selection ratio strategies yielded test information approximately equivalent to the STMI-B 1/1 and MI-B conditions which used no randomization at all.

CONCLUSIONS

This work suggested the following conclusions: (1) The STMI strategy, with Bayesian ability estimation, seems to work about as well as the best adaptive testing strategies; (2) Predictable sequences of test items can be avoided by modifying STMI so that items are selected at random from a nearly optimal set of items; (3) As long as that set is small in number, the adaptive test will not lose an appreciable amount of efficiency; (4) If the set is small to begin with, and gets progressively even smaller (through specification of a shrinking set size) the adaptive test is virtually as efficient as the strategy which chooses the optimal item every time.

REFERENCES

Birnbaum, A. (1968) Some latent-trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Lord, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, New Jersey: Erlbaum.

McKinley, R.L. and Reckase, M.D. (1981) *A comparison of a Bayesian and a maximum likelihood tailored testing procedure* (Research Report 81-2). Columbia, MO: University of Missouri, Educational Psychology Department.

Owen, R.J. (1969) *A Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service.

Owen, R.J. (1975) A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Sympson, J.B., Weiss, D.J. & Ree, M.J. (1982) Predictive validity of conventional and adaptive tests in an Air Force training environment. (Technical Report 81-40) Brooks Air Force Base, Texas: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Vale, C.D. and Weiss, D.J. (1975) *A simulation study of stradaptive ability testing* (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology.

Weiss, D.J. (1974) *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology.

Weiss, D.J. (1982) Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Wetzel, C.D. and McBride, J.R. (1983) *Influence of fallible item parameters on test information during adaptive testing.* (NPRDC TR 83-15). San Diego, CA: Navy Personnel Research and Development Center.

Wood, R.L., Wingersky, M.S., & Lord, F.M. (1976) *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (RM-76-6). Princeton, NJ: Educational Testing Service.

Speeded Tests - Can Computers Improve Measurement?

by
John H. Wolfe
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) contains two speeded tests, Coding Speed(CS) and Numerical Operations (NO). In paper-and pencil mode, these tests are administered with a fixed time limit and scored by counting the number of correct responses. Computerized administration of the same items offers several interesting alternative methods of scoring. For example, one can administer the tests with no time limit, so that everyone finishes the test, and then measure the time each examinee used. Greaud and Green (1985) showed that scoring such a test with a "rate" measure (equal to the number of correct responses divided by the test time) increased overall test reliability. Computer scoring by rates has two advantages over ordinary administration: (1) there is no "ceiling" effect for the fastest examinees, and (2) the scores for the slowest examinees are based on the same number of items as the fastest examinees, and therefore have improved reliability.

Further improvements in reliability can be expected from measuring the examinee's response times for each item, and combining these times into an appropriate total score or scores. As a starting point for proposing alternative scoring functions, consider the Greaud and Green "rate":

$$rate = \frac{number\ correct}{Total\ Time}. \tag{1}$$

By dividing the numerator and denominator by the total number of items, N, the formula is seen to be equivalent to:

$$rate = \frac{P_c}{\frac{1}{N}\sum_{i=1}^{N} T_i} = \frac{P_c}{\bar{T}}, \tag{2}$$

where P_c is the proportion of correct responses and $\bar{T}$ is the sample mean of the item response times, taken over all of the items.

The first method of improving on the formula might be to compute the mean time for only the correct responses. If two examinees take the test, and one of them is able to answer an item wrong twice as fast as the other examinee answers it wrong, it is not clear that the first examinee should be scored higher on the test. It seems plausible that incorrect responses should be eliminated from the scoring.

The second modification to the formula would eliminate "outliers" from the computation of the mean item response time. It is not uncommon for an examinee taking a speeded test to be distracted or pause to ask a question of the proctor. Extraordinarily long response times should be identified and omitted from the scoring.

One potential problem with computing the sample mean of the item times is that the distribution of times is highly skewed. In general, the sample mean of a skewed

distribution is not necessarily a good estimate of the population mean. Thus, a third approach to improving reliability of scoring would be to seek some transformation of the times that makes their distribution more nearly normal. Some possibilities that have been suggested in the literature are $\log T_i$, $\sqrt{T_i}$, and $\frac{1}{T_i}$. To these can be added $\frac{1}{\sqrt{T_i}}$. For each transformation, one can construct a corresponding "rate" measure by computing the mean of the transformed times and then transforming the mean back onto the time scale to get a $\hat{T}$ to replace $\overline{T}$ in equation (2). Thus,

$$rate_{\log T} = \frac{P_c}{\exp\left[\frac{1}{N}\sum_{i=1}^{N}\log T_i\right]}. \tag{3}$$

$$rate_{\frac{1}{T}} = P_c\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\right] \tag{4}$$

$$rate_{\sqrt{T}} = \frac{P_c}{\left[\frac{1}{N}\sum_{i=1}^{N}\sqrt{T_i}\right]^2}. \tag{5}$$

$$rate_{\frac{1}{\sqrt{T}}} = P_c\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{T_i}}\right]^2 \tag{6}$$

METHOD

Eighty-five randomly chosen recruits at the Recruit Training Center, San Diego, were administered two alternate forms of Coding Speed and two alternate forms of Numerical Operations presented on Apple /// personal computers. The order of testing was randomized. The CS tests contained 12 sets of seven items per screen, and the NO tests contained 17 sets of three items per screen. The onset of a new screen was synchronized to the computer so that stimulus timing began when the electron beam was in the upper left corner of the monitor. Timing of key presses was measured with a hardware clock with millisecond accuracy. Pascal software that was used to detect key presses introduced constant errors on the order of a tenth of a second. Each test was administered with a 16-minute time limit, which is more than twice the usual time. This was sufficient for most, but not all subjects to complete the tests.

Outliers were defined as times whose logarithms were more than three standard deviations above the mean log time for that examinee. Reliabilities were computed for all responses, correct responses only, and correct responses trimmed for outliers.

Skewness and kurtosis statistics were computed for individual item times in Form A of the Coding Speed test, along with four transformations of the times. For all tests,

means of the times and their transformations were computed, and five different "rate" scores were obtained. Skewness and alternate form reliabilities were computed for the five means and the five rate scores.

RESULTS

Table 1 shows the alternate form reliabilities of the Greaud and Green rate scores when all item times are scored, when only correct responses are scored, and when correct responses are trimmed at the upper tail for outliers. Restricting the scoring to correct responses appeared to have no effect on reliability, but trimming outliers raised reliability somewhat.

Table 1
Effect of Trimming Incorrect Responses and Outliers On Reliability of Rate Scores

	Coding Speed	Numerical Operations
All Responses	.81	.72
Correct Responses	.80	.73
Trimmed Correct Responses	.83	.75

Table 2 summarizes the skewness and kurtosis characteristics of 84 Form A Coding Speed item times and their transformations. As expected, the times were quite skewed, and in addition, had considerable kurtosis. Taking logs of the times eliminated the skewness, and also reduced kurtosis. The reciprocal transformation made matters much worse. The square root transformation reduced skewness, but not as much as the logarithms. Taking the reciprocal of the square roots made skewness and kurtosis worse. From these data, it appears that the log transformation is best, and that the

reciprocal should not be used to normalize the data.

Table 2
Mean Skewness and Kurtosis of 84 Coding Speed Item Times and Their Transformations

	Skewness	Kurtosis
T_i	0.99	1.47
$\log T_i$	-0.03	0.84
$1/T_i$	1.27	5.50
$\sqrt{T_i}$	0.51	0.62
$1/\sqrt{T_i}$	0.64	2.55

Table 3 shows the skewness of five different means and five "rates" based on these means. Although the central limit theorem merely implies that the means of the times should be normally distributed within each individual, it is still somewhat surprising that the mean times, reciprocals, and square roots are significantly skewed across individuals. Again, the log transformation showed the least skewness. None of the rate measures were skewed in the Coding Speed tests, and all of the rates were significantly skewed in the Numerical Operations tests.

Table 3
Skewness of Alternative Scoring Formulas

	Coding Speed		Numerical Operations	
	Form A	Form B	Form A	Form B
Mean T_i	0.84*	1.46*	0.93*	0.52*
Mean $\log T_i$	0.14	0.31	0.16	-0.15
Mean $1/T_i$	0.64*	0.41	0.56*	0.77*
Mean $\sqrt{T_i}$	0.48	0.81*	0.54*	0.17
Mean $1/\sqrt{T_i}$	0.21	0.08	0.20	0.46
$P_c/(Mean\ T_i)$	-0.18	0.07	0.52*	0.61*
$P_c/Exp(Mean \log T_i)$	-0.13	0.05	0.56*	0.66*
$P_c(Mean\ 1/T_i)$	-0.00	0.03	0.60*	0.72*
$P_c/(Mean \sqrt{T_i})^2$	-0.17	0.06	0.54*	0.64*
$P_c(Mean\ 1/\sqrt{T_i})^2$	-0.08	0.04	0.58*	0.69*

* Significant at $p < .05$ by 2-tailed *t*-test.

Table 4 is an expanded version of Table 1, in which mean times, mean log times, and rates based on mean log times are also shown. For all measures, it appears that restricting scoring to correct responses did not improve reliability. Eliminating outliers did improve reliability. The log transformation improved reliability if outliers were not trimmed, but not otherwise. Trimming outliers improved reliability if untransformed times were used, but not otherwise. Rate scores were more reliable than average times for Coding Speed, but not for Numerical Operations.

Table 4
Reliabilities of Alternative Scoring Functions

	Coding Speed	Numerical Operations
	All Responses	
Mean Time	.76	.74
Mean Log Time	.79	.75
P_c/Mean Time	.81	.72
P_c/Exp(Mean Log Time)	.82	.75
	Untrimmed Correct Responses	
Mean Time	.74	.74
Mean Log Time	.78	.76
P_c/Mean Time	.80	.73
P_c/Exp(Mean Log Time)	.81	.75
	Trimmed Correct Responses	
Mean Time	.77	.76
Mean Log Time	.78	.76
P_c/Mean Time	.83	.75
P_c/Exp(Mean Log Time)	.82	.76

DISCUSSION

The results presented in this paper are only preliminary: more subjects remain to be tested, and additional analyses need to be performed, particularly on the Numerical Operations items. Nevertheless, certain conclusions and directions for future work stand out:

One successful method for improving reliability is to eliminate outliers. Future work should explore this method in more detail. The optimal cutting point for trimming the data is a question that needs to be answered empirically.

Another method that improved reliability as much as trimming outliers was the log transformation of times. So far, there is no indication that both methods combined

are better than one of them. However, combining both methods does not decrease reliability, and may inspire greater confidence. In the end, considerations of computational speed and program complexity may be the determining factors in deciding which method(s) to use.

In all of the work described here, it has been implicitly assumed that the items are homogeneous in difficulty, and only individual differences have been examined. The next step in the research should be to examine items, and to develop a model that encompasses both item characteristics and individual differences in ability. This approach should be especially useful in Numerical Operations, where addition, subtraction, multiplication, and division have quite different average response times.

REFERENCE NOTES

Greaud, V.A. & Green, B.G. (1985) *Equivalence of conventional and computer presentation of speed tests*. Unpublished manuscript.

MEDIUM OF ADMINISTRATION EFFECTS ON ATTITUDES TOWARD ASVAB TESTING

William F. Kieckhaefer
RGI, Incorporated

Daniel O. Segall
Kathleen E. Moreno
Navy Personnel R&D Center

BACKGROUND

Computerized adaptive testing is being considered for use in military selection and classification. One research question under investigation concerns whether medium of administration affects examinee attitudes toward the current Armed Services Vocational Aptitude Battery (P&P-ASVAB). Recognizing the importance of this, the Defense Advisory Committee on Military Personnel Testing recommended that the reactions of examinees to the computerized adaptive version of that test (CAT-ASVAB) should be systematically collected and analyzed (Linn, Bond, Britell, Campbell, Jaeger, Novick, and Uhlaner, 1983).

Early researchers reported on particular aspects of the relation between attitudes and computerized testing. For example, Hedl, O'Neil, and Hansen (1973) showed that less favorable reactions of the subjects to a computerized test were due to a lack of clarity in the instructions and unfamiliarity with computer terminals. In a related study, Walther and O'Neil (1974) found that subjects with greater test anxiety or negative attitudes toward computers performed more slowly and made more errors on a test than subjects with lower levels of test anxiety. More recently, Nilles, Carlson, Gray, Hayes, Holmen, and White (1980) supported the view that computer usage generates anxiety which negatively impacts on user attitudes and performance.

Other researchers present a different view. Schmidt, Urry, and Gugel (1978) found that examinees completing a computerized adaptive test had positive attitudes toward it. In a study on computer-managed instruction, Robinson, Tomblin, and Houston (1982) also reported positive user attitudes.

More specifically for ASVAB testing, Mitchell, Hardwicke, Segall, and Vicino (1983) reported generally positive attitudes of male Navy recruits toward taking the CAT-ASVAB. Some attitudes corresponded with subjects' level of experience with computers or keyboards. For example, those with "little to none" computer experience were more likely to indicate that computerized testing was more impersonal than paper-and-pencil testing. Subjects with "little to none" keyboard experience were generally more likely to express uneasiness about taking a test on a computer and to indicate that the computerized test was more difficult. This research showed that while computerized testing may be considered impersonal by some, this perception does not imply a negative attitude toward computerized tests. These findings were supported in the service-wide study of Hardwicke and Yoes (1984).

PURPOSE

Investigations of user attitudes toward computerized testing can benefit from clearer attitudinal items and improved experimental designs. Furthermore, little is known about the relation between attitudes toward ASVAB testing and examinee ability or such background variables as race and sex. The purpose of this study was to assess the effects of the medium of administration on attitudes toward ASVAB testing. While some research does exist in this area, this research is still exploratory. Therefore, the research objectives were to:

(1) Examine the effect of medium of administration on examinee attitudes toward testing; and

89

(2) Determine if this relation differs as a function of sex, race, or ability.

METHOD

This is part of a larger study investigating several aspects of subgroup differences in ASVAB testing. While a total of 3,094 Navy recruits were tested under the larger study, only the data for 619 recruits tested at the Recruit Training Center in Orlando, Florida were available at the time of this presentation. Therefore, findings presented here are preliminary. The sample was 60.9% male and 39.1% female. Also, 49.4% were White, 18.7% were Hispanic, and 31.7% were Black.

Test proctors selected all recruit companies of men and women which began training between April and June of 1985. Then proctors randomly selected recruits from a company for testing and randomly assigned them to complete either the P&P-ASVAB or the CAT-ASVAB first. These were the Medium of Administration conditions. Test proctors collected the background information regarding examinee sex and race from each subject's enlistment form (DD form 1966). They also obtained Pre-enlistment scores on the Armed Forces Qualification Test (AFQT) from that form.

After completing the first version of the ASVAB (i.e., P&P or CAT), subjects responded on a seven-point scale to nine questions presented in a paper-and-pencil format. These were the dependent variables in the present study. While recruits did complete the other ASVAB version also, data from the second version are not relevant to this attitude study and are not reported here.

RESULTS

Responses to each of the nine attitude questions served as the dependent variables in a three-factor analysis of variance: Medium by Sex by Race. The significance of the AFQT covariate was obtained for each dependent variable. An analysis of covariance was performed for items with significant AFQT covariates.

For each attitude question, Table 1 shows the number of respondents (N), the overall mean, and the cell means for significant main effects in the analysis of variance. Figure 1 depicts the significant interaction effects involving medium. Table 2 shows each attitude question and the distribution of responses. For significant medium effects, the distribution is shown for each group.

Only question 4 had a significant-AFQT covariate. When entered into an analysis of covariance, two of the previously significant effects were no longer significant: the main effect of Race and the interaction effect of Sex by Race.

DISCUSSION

While these are preliminary findings, six of the nine attitude questions had significant main effects of Medium--all favoring the CAT-ASVAB. Those taking the CAT-ASVAB felt better about taking the test battery. In addition, they felt less tired, they experienced less eye strain, and they thought the test battery was shorter. These results are probably due to the fewer number of items required by the CAT-ASVAB. In general, the CAT-ASVAB measures the same content areas as the P&P-ASVAB with about half the number of items.

Furthermore, those taking the CAT-ASVAB felt more relaxed during the test. Perhaps this is because they proceeded at their own pace while those taking the P&P-ASVAB had specified time limits. Also, those taking the CAT-ASVAB reported that the instructions were clearer. This may be attributable to the interactive nature of the instructions, which included immediate feedback during a keyboard familiarization sequence and sample questions.

Three questions showed no main effects of medium. These indicated no differences between administration conditions on anxiety, perceived difficulty of the questions, and perceived fairness.

Table 1
Cell Means for Main Effects

Item	Attitude	N	Overall	Medium[a]	Sex[b]	Race[c]
1	Overall feelings	597	2.7	2.4 3.0		
2	Fatigue	596	3.5	4.0 3.1		
3	Anxiety	595	3.5			
4	Question Difficulty	594	3.9		3.8 4.0	4.0[d] 3.8 3.8
5	Fairness	595	2.2			
6	Test Length	593	4.0	3.4 4.5		
7	Pressure	593	3.0	2.7 3.2		
8	Eye Fatigue	593	3.3	3.4 3.2		3.5 3.2 3.2
9	Instruction Clarity	593	1.5	1.4 1.6	1.6 1.4	

[a] The order of cell means is CAT-ASVAB, P&P-ASVAB.
[b] The order of cell means is Male, Female.
[c] The order of cell means is Black, Hispanic, White.
[d] This effect was not significant when AFQT was a covariate.

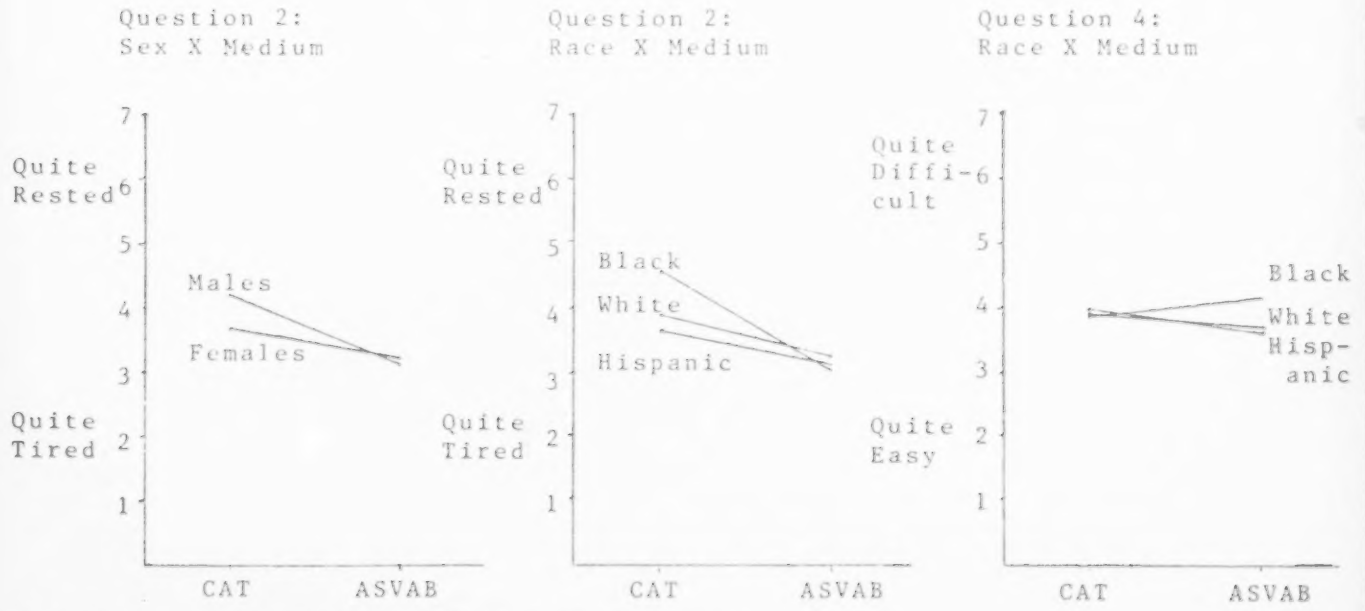


Figure 1. Interaction effects with Medium.

Table 2

Distribution of Responses to Attitude Questions

1. Overall, how did you feel about taking the test battery?

	1 *extremely good*	2 *quite good*	3 *slightly good*	4 *neither*	5 *slightly bad*	6 *quite bad*	7 *extremely bad*
CAT-ASVAB:	16.4%	42.7%	26.3%	9.5%	4.7%	0.4%	0.0%
P&P-ASVAB:	5.6%	35.6%	26.9%	24.1%	5.3%	2.2%	0.3%

2. Overall, how tired did you feel at the end of the test?

	1 *extremely tired*	2 *quite tired*	3 *slightly tired*	4 *neither*	5 *slightly rested*	6 *quite rested*	7 *extremely rested*
CAT-ASVAB:	2.2%	8.4%	35.0%	22.6%	10.2%	16.8%	4.7%
P&P ASVAB:	5.3%	19.9%	47.2%	18.0%	4.3%	4.3%	0.9%

3. During the test, how anxious did you feel?

	1 *extremely calm*	2 *quite calm*	3 *slightly calm*	4 *neither*	5 *slightly anxious*	6 *quite anxious*	7 *extremely anxious*
Overall:[a]	8.4%	29.6%	11.6%	17.5%	25.0%	6.2%	1.7%

4. What is your opinion of the difficulty of the questions?

	1 *extremely easy*	2 *quite easy*	3 *slightly easy*	4 *neither*	5 *slightly difficult*	6 *quite difficult*	7 *extremely difficult*
Overall:[a]	1.0%	14.5%	20.4%	28.8%	31.8%	3.5%	0.0%

5. How fair do you feel the test was?

	1 *extremely fair*	2 *quite fair*	3 *slightly fair*	4 *neither*	5 *slightly unfair*	6 *quite unfair*	7 *extremely unfair*
Overall:[a]	23.9%	55.1%	7.7%	9.6%	2.4%	0.8%	0.5%

[a] For comparisons with non-significant medium effects, the distribution of responses is shown for the combined CAT-ASVAB and P&P-ASVAB groups

Table 2

(continued)

6. What is your opinion of the length of the test battery?

	1 *extremely short*	2 *quite short*	3 *slightly short*	4 *neither*	5 *slightly long*	6 *quite long*	7 *extremely long*
CAT-ASVAB:	6.6%	17.9%	21.2%	38.3%	13.9%	1.8%	0.4%
P&P-ASVAB:	0.9%	5.0%	10.0%	33.8%	30.0%	13.8%	6.6%

7. During the test, how relaxed or pressured did you feel?

	1 *extremely relaxed*	2 *quite relaxed*	3 *slightly relaxed*	4 *neither*	5 *slightly pressured*	6 *quite pressured*	7 *extremely pressured*
CAT-ASVAB:	21.2%	35.0%	17.2%	12.0%	12.8%	1.8%	0.0%
P&P-ASVAB:	10.3%	29.8%	17.6%	16.9%	20.7%	3.8%	0.9%

8. During the test, how strained or tired did your eyes feel?

	1 *extremely tired*	2 *quite tired*	3 *slightly tired*	4 *neither*	5 *slightly rested*	6 *quite rested*	7 *extremely rested*
CAT-ASVAB:	7.7%	17.2%	37.6%	17.5%	5.8%	9.1%	5.1%
P&P-ASVAB:	8.2%	22.9%	33.9%	21.0%	5.3%	6.9%	1.9%

9. How clear do you feel the instructions were?

	1 *extremely clear*	2 *quite clear*	3 *slightly clear*	4 *neither*	5 *slightly confusing*	6 *quite confusing*	7 *extremely confusing*
CAT-ASVAB:	68.2%	29.2%	0.7%	0.7%	1.1%	0.0%	0.0%
P&P-ASVAB:	53.0%	41.7%	1.6%	1.6%	0.9%	0.6%	0.6%

The other main and interaction effects are not so readily interpreted. A cultural bias hypothesis would favor White males over all other subgroups. Since this result occurred for only four of the remaining nine main and interaction effects, these results do not support a cultural bias hypothesis.

When all the data are ready, analyses will include measures of computer knowledge and test performance. Then, studies will investigate the relations between attitudes, computer knowledge, biographical characteristics, ability, and test performance.

References

Hardwicke, S. B., & Yoes, M. E. (1984). Attitudes and performance on computerized vs. paper-and-pencil tests. Paper presented at the Second Annual Air Force Conference on Technology in Training and Education, Wichita Falls, Texas.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology, 40,* 217-222.

[illegible] R. L., Bond, L., Britell, J. K., Campbell, J. P., Jaeger, R. M, Novick, M. R., & Uhlaner, [illegible] (1983). *Report of the March 3 & 4, 1983 Meeting of the Defense Advisory Committee on Military Personnel Testing.* San Diego, CA.

[illegible]ell, P. A., Hardwicke, S. B., Segall, D. O., & Vicino, F. L. (1983). Computerized adaptive testing: A preliminary study of user acceptability. *Proceedings of the 25th Annual Conference of the Military Testing Association.*

Nilles, J., Carlson, F. R., Gray, P., Hayes. J., Holmen, M., & White, M. J. (1980). *Technology assessment of personal computers.* Los Angeles, CA: University of Southern California, Center for Future Research.

Robinson, C. A., Tomblin, E. A., & Houston, A. (1981). *Computer-managed instruction in Navy technical training: An attitudinal survey* (Report No. 82-19). San Diego, CA: Navy Personnel research and Development Center.

Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer-assisted tailored testing: Examinees reactions and evaluations. *Educational and Psychological Measurement, 38,* 265-273.

Walther, G. H., & O'Neil, H. F. Jr. (1974). On-line user-computer interface--The effects of interface flexibility, terminal type, and experience on performance. *AFIPS Conference Proceedings, 43,* 379-384.

STATISTICAL PROCESS CONTROLS AS AN ENHANCEMENT TO JOB DESIGN

Steven L. Dockstader
Navy Personnel Research and Development Center

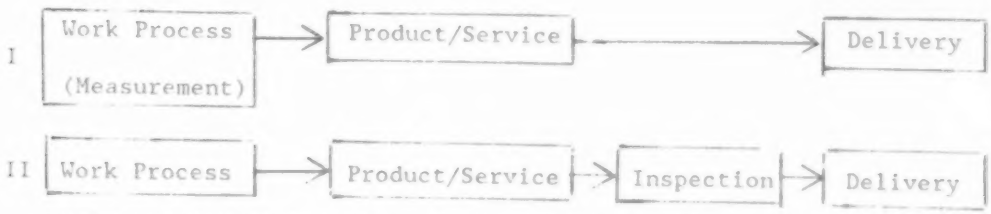
Abstract

The paper presents an argument for adopting statistical process control as an approach for improving productivity in organizations. The fundamental problems of resitance to change are discussed and job changes resulting from SPC are logically analyzed in terms of job characteristics theory. It is concluded that performance increases can be expected from the approach, but whether they are on the ability or motivational dimension is unclear and should be the subject of future research

Introduction

The purpose of this paper is to attempt to provide an explanation for the varying degrees of success found for projects seeking to improve productivity through changes in quality controls. It is the contention of this author that productivity derived from quality improvement is the result of (a) leadership initiatives which support a change in the methods of quality control and (b) factors intrinsic to the methods which contribute to both their acceptance and incentive value to the individual.

There are two basic approaches used to achieve product or service quality. The methods vary on the extent to which the responsibility for quality is vested in the producer of the product or service or with an external agent, such as a quality checker. The methods can be described schematically as follows:



Approach I will be referred to as the process control approach, while approach II will be referred to as the product inspection approach. Proponents of process control contend that it leads to higher levels of quality and lower costs than the product inspection approach. Sherkenbach (1984) has argued that greater efficiencies are achieved by the process control approach because the methods, over time, eliminate the problems in the (production) system which result in defects. This, in turn eliminates waste, rework, and the need for inspectors. Juran (1974), in a massive review of process control work, has pointed out that systems problems account for the vast majority of quality defects--as high as 85%--and that the systems problems can best be dealt with by the process control approach.

Problem

While the arguments of Juran and Sherkenbach have both logical and empirical support--most dramatically in Japanese industry--the question remains as to why such a successful approach has not achieved wide acceptance within the United States where the methods were initially developed. Tribus (1984) has argued that the management philosophy of most American business and industry is such that there is resistance to turn over control systems to the workforce, i.e., that quality is properly the responsibility of management and can not and should not be deployed to the workforce. He then goes on to point out that there has been no economic reason to change from the current system because U.S. manufactured goods have dominated world markets since WWII and costs associated with the product inspection approach to quality control were of no consequence in a noncompetitive market.

Recent change in world market conditions have forced many companies to re-examine their approach to quality (Houston et al.*, 1985). Being less driven by profits and competition, public sector organizations can be expected to resist the change to the more effective process control approach. From this, it would appear that a substantial amount of education of management and concerted action by leadership must occur. This argument has been made most strongly by Deming (1985), although it applies to almost all types of organizational change efforts of this magnitude (see Metz, 1984 for an excellent discussion). Part of the education of management should concern an understanding of the resistances to change in control systems and the likelihood that they can be successfully overcome. The balance of this paper will address these issues.

Factors Influencing Acceptance of Control Systems

Lawler (1976) has analyzed the nature of control systems in organizations and the kinds of resistances to their development. In his analysis, he points out that resistance to control systems is most likely when:

1. The control system measures performance in a new area.
2. The control system replaces a system that people have a high investment in maintaining.
3. The standards for control are set without participation.
4. Control system feedback does not go to those who are measured.
5. Control system feedback goes to higher levels in the organization and is used in the reward system.
6. People affected by the existing system are relatively satisfied and see themselves committed to the organization.
7. The people affected by the system are low in self-esteem and authoritarianism.

Using Lawler's analysis, we can examine the resistance to change from the product inspection approach to the process control approach. We can also use the information to suggest ways in which leaders can approach the change process. Following that, an argument will be made on the basis of contemporary work motivation theory as to why the process control approach should be very successful, once implemented.

In the following discussion, Lawler's seven sources of resistance will be considered in turn for the operations of a large aircraft overhaul facility in the Navy. Although the observations made here are not based upon empirical data, they have been corroborated by senior managers in that organization and by on-site research personnel.

1. The process control approach does measure performance in new areas. In fact, the essence of the approach involves measuring several significant features about the production process prior to completion of a product or service.

2. Ultimately many of the personnel who are currently used in the quality control department will be deployed to other parts of the organization, or be conducting quality control activities not currently being performed (e.g., incoming supplies, customer services, etc.). This displacement and/or retraining of personnel is viewed as a threat by those in the current quality control function.

3. Quality control standards are usually established by engineers or quality technicians. In the case of process control, however, a fixed standard has no meaning. Control is defined by taking actions to keep the process within variability limits which are determined by the process itself. Because the limits change as a function of improvements in the system, no fixed standard can be applied.

4. Using the process control approach, the basic data is collected by the performer. In this sense, feedback is immediate. Furthermore, because the information gathered is typically a historical record with relational information on the record (e.g., a control chart), the worker can evaluate the data and determine what action, if any, need be taken.

5. Whether the data is fed to higher levels and used within the reward system depends upon a number of factors. The most significant is the degree to which the worker has discretion to make decisions concerning corrections to the system. This, in turn, is usually based upon the extent of the system changes and their costs, but could also be a reflection of the management philosophy of the organization. This will be considered in greater detail in a subsequent discussion.

6. This is the "status quo" factor, and it can be said that a change in the intertial state of the organization will be determined by whether or not a "critical mass" (Deming, 1985b) can be developed to overcome the status quo. The state of inertia in most bureaucracies, such as those in most large bureaucratic organizations, is at steady state and resistant to change under normal workload conditions.

7. It is difficult to assess this factor. The people most affected by the quality control system are those in the "production" area. As a group, they are the largest in number and exert the greatest influence on achieving the mission of the organization. However, under the current product inspection approach, they receive the most censure when product quality does not meet specifications/test. Managers have been of the opinion that this has led the workers to lose identity with the quality of their products because someone else has been responsible for detecting it.

Lawler has indicated that, to the extent that these factors hold for workers, they will engage in non-productive or even counter productive behaviors. Using his analysis and the previous discussion, it appears that the process control approach should meet less resistance in terms of factors 3, 4, 6, and 7. That is to say that it (a) does not deal with standards per se, (b) provides feedback information to the performer, (c) is of greater benefit to most of the workforce than the existing system and (d) can enhance the self esteem of the worker as he begins to take charge of the quality of his work.

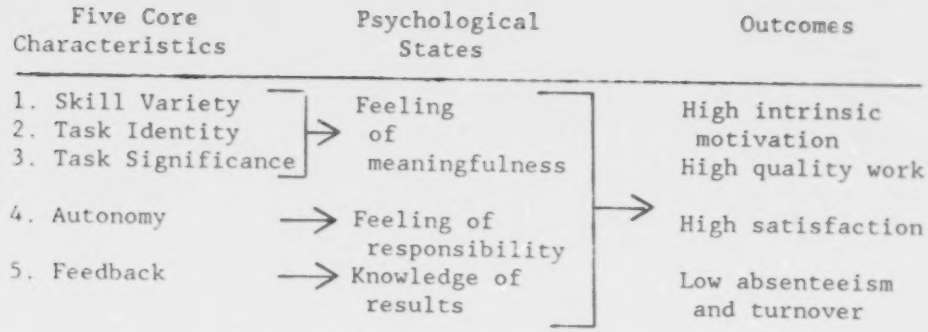
Of the other factors, only the second appears to be of significant concern in terms of resistance to change. In the organization under study, quality control is vested in a functional department. While performing inspections or audits of the work conducted in the production area is not their only function, it does define their central raison d'etre. In addition, this organization is one of several which reports to a headquarters. Both the headquarters and the sister organizations contain quality control functions based upon inspection and audit. Resistance here would have to be overcome.

Factors 1 and 5 are potentially areas of resistance because of the new measures and added work required (1) and because the information could be used to evaluate performance of the workers (5). Neither of these is necessarily negative, but the workforce is often wary concerning the use of performance measures. If the management philosophy and the culture of the organization is one that rewards improvement then there will be little resistance.

Process Control and Worker Motivation

Our discussion thus far has focussed on the desirability of changing from product inspection to process control and the nature of resistances in making such a change. While it appears obvious that such a change is both desireable and feasible from a management standpoint, what is in it for the worker? After all, with the exception of some of the existing quality control personnel, the major job changes will be that of the worker and perhaps his immediate supervisor. If this change is not seen by the worker as having incentive value, then it will very likely be resisted.

Job Characteristics Theory (Hackman & Lawler, 1971; Hackman & Oldham, 1976) provides a conceptual framework to evaluate the design of a workers job to include the process control approach to quality control. The theory is based upon a plethora of research which has revealed that there are three psychological states which contribute to worker motivation. These are feelings of meaningfulness, responsibility, and knowledge of results. The theory goes on to describe what job characteristics will result in these feelings. The theoretical relationships can be schematicized as follows:



Considering each of these characteristics in turn, we can determine the motivating potential, or incentive, of the workers job when process control becomes a part of the job. Skill variety is obviously increased because the job will now involve collection of data, charting of data, and reporting process aberrations. Task identity should increase because attention will be focussed on aspects of the process which were previously receiving less formal, e.g., measurement, attention. The perceived significance of the task may also be enhanced because taking process control actions should occasion interaction with supervisors, staff, and managers which would not ordinarily occur. Autonomy will be increased because quality control actions and responsibilities will now be formally placed in the hands of the worker. Finally, feedback will be immediate in terms of the things being measured. Feedback as a result of process changes will, in most cases, be immediate as well.

From this logical analysis and the model displayed in Table 1, we can predict the outcomes displayed there. During the course of the forthcoming year, these hypothetical relationships will be tested in the Navy maintenance environment. The use of process control as a method to enrich jobs has not received attention in the empirical literature, but the aforementioned analysis suggests that it should be an effective way to motivate workers as well as increase the quality of their efforts.

Locke (1980) has indicated that job enrichment has not been effective in motivating employee performance, when the effects of goal setting have been controlled. Such a confounding of variables is not expected in the process control situation, because the demand characteristics of process control are not upon employee effort, but upon removal of system-generated variation. Performance increases, then, would probably result from enhanced ability to perform the job, as opposed to the motivation to increase effort.

This latter point begs the question of whether employee motivation could be at all affected by adopting the process control approach. Lawler's (1976) discussion of expectancy models of worker motivation would suggest that, in the absence of extrinsic rewards for quality improvement, the motivational impact can only be derived from (a) the E→P probability or (b) or the valence associated with outcomes other than extrinsic rewards. From our logical analysis above, it is not clear which of the two, or some combination of both could account for changes in the motivating potential of a job enriched by inclusion of process controls for quality. These questions will be addressed in future research.

References

Deming, W. E. Transformation of western style of management. Interfaces 15: 3 May-June 1985 (pp. 6-11)

Hackman, J. R. & Lawler, E. E. Employee reactions to job characteristics. Journal of Applied Psychology, 55, 1971, 259-286.

Hackman, J. R. & Oldham, G. R. Development of the Job Diagnostic Survey. Journal of Applied Psychology, 60, 1975, 159-170.

Houston, A., Shettel-Neuber, J., & Sheposh, J. Management methods for quality improvement based on process analysis and control. NPRDC Technical Report, 1985 (in press).

Juran, J. Quality Control Handbook. New York: McGraw Hill, 1974.

Lawler, E. E. Control systems in organizations. In Dunnette, M. D. (Ed.) Handbook of Industrial and Organizational Psychology, Chicago: Rand McNally, 1976.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N. & Denny, A. T. The relative effectiveness of four methods of motivating employee performance. In Duncan (Ed.) Changes In Working Life, Wiley. 1980.

Metz, E. J. Managing change: Implementing productivity and quality improvements. National Productivity Review, Summer 1984, 303-314.

Scherkenbach, W. W. The process of continuous improvement. Presented at the symposium "Managing Systems of People and Machines for Quality and Productivity", San Diego, 1984.

Tribus, M. Creating the quality company. Presented at the symposium "Managing Systems of People and Machines for Quality and Productivity", San Diego, 1984.

A GROUP WAGE INCENTIVE SYSTEM: DESIGN AND IMPLEMENTATION ISSUES

Deborah A. Mohr
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

In an effort to improve performance and reduce costs in a naval shipyard, a group wage incentive system for production workers was developed, implemented, and evaluated. The system was designed to improve performance efficiency without negatively affecting schedule adherence, product quality, or workers' job attitudes.

This project was part of a continuing research program to investigate the effects of wage incentive systems in Navy industrial facilities. Previous projects evaluated the effects of performance contingent reward systems (PCRSs) with a variety of civil service employees: key entry operators, small purchase buyers, and aircraft engine mechanics. Under a PCRS, employees earn cash bonuses (incentive awards) for work performed above established standards. The more performance exceeds the standard, the larger the bonus. PCRS rewards are paid through existing award programs, are recurrent (being accrued as often as performance exceeds standard), and are in addition to employees' base salary.

The present effort differed from previous projects in that awards were based on measures of group performance. Shipyard production workers typically work together in teams (called work gangs) of 10 to 20 employees supervised by one foreman. Thus, a PCRS based on measures of group performance was more appropriate than one based on individual performance measures.

INCENTIVE SYSTEM OVERVIEW

The performance measure used for this system was one of performance efficiency. It was calculated by dividing the manhours allowed to complete a work gang's jobs by the manhours actually expended to complete the work. Thus, when work is completed in exactly the time allowed for that work, the work gang's performance efficiency, called a performance factor (PF), is 100 percent. When work is completed in less time than the allowance, the gang's PF will be greater than 100 percent and manhours will be saved. Both inputs to this measure (manhours allowed and expended) were routinely collected by the shipyard's management information system (MIS). Prior to implementation, the shipyard MIS was further enhanced to provide more accurate performance measures and to provide monthly automated incentive award calculations and continual award tracking.

Under the shipyard's PCRS, work gangs were eligible for awards whenever they saved manhours by completing their jobs in less time than the manhours allowed for those jobs. The value of these saved hours was shared with employees in the form of incentive awards. The work gang's saved hours were distributed to members based on each worker's contribution to the workgang (his or her share of the gang's total work hours). Based on the 50 percent sharing rate used during the system test, half of the cost savings associated with a work gang's manhour savings were paid out to gang members as incentive awards. The remaining 50 percent was retained by the shipyard. The actual value of each saved hour was based on the employee's accelerated hourly wage rate.

A similar incentive system was established for shop foremen in which all foremen comprised one group eligible for awards whenever performance of the entire shop resulted in manhour savings. In addition, to encourage foremen to work together to help the shop improve, each foreman received a one-time bonus of $125 the first time the shop's PF exceeded 100 percent.

Because the shop selected for test of the incentive system historically spent many more manhours to complete its work than were allowed, few work gangs would save manhours and earn incentives at typical performance levels. Since incentive systems do not motivate employees to improve performance unless they believe it's possible to earn awards, shipyard managers decided to adjust all performance measures upward by 10 percent for the purposes of subsequent award calculations. Thus, work gangs actually accrued manhour savings whenever their PF exceeded 90 percent and foreman earned awards whenever the shop's PF exceeded 90 percent. Despite this adjustment, the incentive system rewarded employees for performance improvement.

IMPLEMENTATION

Prior to implementation of the test system, a shipyard instruction was issued documenting the incentive system and specifying responsibilities during the test period. A senior military officer was assigned as project manager and a general foremen within the test shop served as system coordinator. An agreement was negotiated with the local union and approval was obtained from the appropriate headquarters commands. Finally, employees and supervisors in the test shop were given training to assure their understanding of the enhanced performance measurement system and the group incentive system. The PCRS was then implemented for test in Shop 31, the shipyard's inside machine shop. Shop 31 is one of 17 shops at the shipyard and employs approximately 480 wage grade employees and 23 foremen assigned to 18 work gangs.

RESULTS

During the 19 months of the system test a total of $177,000 was earned by employees. Sixteen of the eighteen work gangs (comprising 89 percent of shop employees) earned awards. Of those employees earning awards, the average total earnings during the test period was $419. (Total earnings ranged from $1 to $2488.) Foremen earnings averaged $237 for the 2 months that the shop's performance factor exceeded 90 percent.

Evaluation of the incentive system test revealed that the program produced a significant increase in the shop's performance efficiency (see Figure 1). For analysis purposes the 19 month test was divided into two phases: Incentive Phase 1 consisted of the first 8 4-week incentive periods and Incentive Phase 2 consisted of the remaining 11 4-week incentive periods. The shop showed a 7.5 percent improvement over average baseline performance during Incentive Phase 2 (an improvement from 91.4% to 97.6%). During the first incentive phase, the shop maintained its baseline performance despite a severe workload reduction (performance averaged 89.0%). Two comparison shops (see Table 1) showed substantial performance decreases during the same time, although their workload reductions were less severe than that of the test shop. Since the end of the 19-month test, the test shop has maintained its improved performance. As expected, implementation of the system did not cause any negative effects on the shop's schedule adherence or product quality.

Participants' job attitudes and evaluations of the incentive system were assessed during the test. Although recognizing certain problems related to system operation (particularly the effects of the workload reduction), 80% of those expressing an opinion favored continuing the incentive system. No positive or negative effects on workers' job attitudes (e.g., job satisfaction and job stress) were found.

A cost savings analysis revealed that the net cost savings due to improvements over baseline performance during the system test exceeded $600,000. If similar results occurred following expansion of the system to all other production shops, the shipyard could realize net cost savings of approximately $6,794,000 annually.

The shipyard realized a number of concurrent positive outcomes from the system test, including improvements in shop practices and initiation of management actions directed toward resolving productivity impediments. Foremen began taking greater care in preparing employee time cards, correcting labor mischarges, and reviewing work documents before beginning jobs. As a result of the interest in improvement engendered by the incentive system, a number of productivity impediments were highlighted during the test. These impediments were not new. Rather, they were long-standing shipyard problems that became

more salient when money was tied to performance. The incentive system provided the impetus to attack these problems and as a result a shipyard-wide problem-solving team was established and successfully resolved a number of these issues.

IMPLEMENTATION AND MAINTENANCE ISSUES

Throughout this effort, various issues surfaced that revealed the complexity of designing and implementing productivity improvement systems in real organizations. While the success of the test system indicates these issues can be effectively resolved, nonetheless, some important conclusions can be drawn from this project.

Several issues arose during the design of the incentive system. The decision to develop a group system was a logical result of analysis of the shipyard's work settings. Implementation of a typical incentive system (most likely based on measures on individual performance) would have been inappropriate and quite possibly ineffective. Managers considering a PCRS should realize that no standard system exists. The PCRS must be designed to fit the organization and its priorities.

A number of incentive system parameters had to be specified in the design phase, as well. These included: the incentive level (the performance level at which employees were eligible to earn awards), the sharing rate (the proportion of cost savings shared with employees), and the savings distribution method (the way savings were shared among work gang members). As discussed, the incentive level was dropped to 90 percent in the belief that this level would be seen as attainable by workers and that it would improve the motivating potential of the system. The 50 percent sharing rate used in this test was selected because it was the maximum allowable by federal regulations and because it was likely to be perceived as fair to both employees and the organization. Distribution of savings based on worker inputs was used to further strengthen participants' perceptions of fairness. These design parameters can be assumed to have been effective based on the favorable results of the test period. However, there is no way to determine if different parameters might have been substantially more effective. Little research has been done to investigate the effectiveness of different levels of these parameters (e.g., a 20% vs. a 50% sharing rate) or the situational variables that require parameter changes.

During the implementation and maintenance phases, additional issues arose. Primary among these was management's relationship to the incentive system. The success of organizational change efforts such as incentive systems is at least in part contingent on active support from management. A high degree of commitment to the program is necessary before and after implementation, commitment involving more than just verbal

support. It is difficult to implement effective changes when either top management or those expected to implement change are unsupportive.

During the test of the incentive system, the shipyard experienced a rather significant workload reduction. This appeared to have precluded performance improvement until the shop's staffing level was brought into balance. Workers are unlikely to improve their productivity if, by so doing, they believe they will risk their jobs. In implementing performance improvement programs, managers must continually address the balance between workload and staffing within the organization and must develop means to capitalize on the effects of resulting improvement.

Tying money to performance highlighted a number of long-standing shipyard problems that were subsequently addressed by a problem-solving team. The importance of tapping this increased interest in performance improvement cannot be overestimated. Many shipyard managers believed that the incentive system's major benefits were in encouraging supervisors to do their jobs and in focusing efforts on resolving productivity impediments. Such auxiliary benefits of incentive systems should not be overlooked.

Finally, the issue of incentive system expansion surfaced. To continue to run a test system in only one shop is not feasible. With proven cost savings resulting from performance improvement the next logical step is to expand to other shops. Managers must carefully consider how and how far to expand a successful incentive system. During expansion, care must be taken to adapt the system to other sites and to continue to monitor its effectiveness. Managers should also consider means to include production support and other indirect workers in such incentive systems.

Although there are a substantial number of complex issues that must be faced in developing and implementing wage incentive systems, their proven effectiveness indicates that such efforts are worthwhile. Further research to identify effective design parameters, the increased use of automation to support wage incentive systems, and the benefits that can be derived from additional experience with these systems will help to limit the effort required to design and implement wage incentive systems in the future.

The opinions expressed in this paper are those of the author and should not be construed as official or as reflecting the views of the Department of the Navy.

Figure 1

Trends in Performance Factors for Shop 31

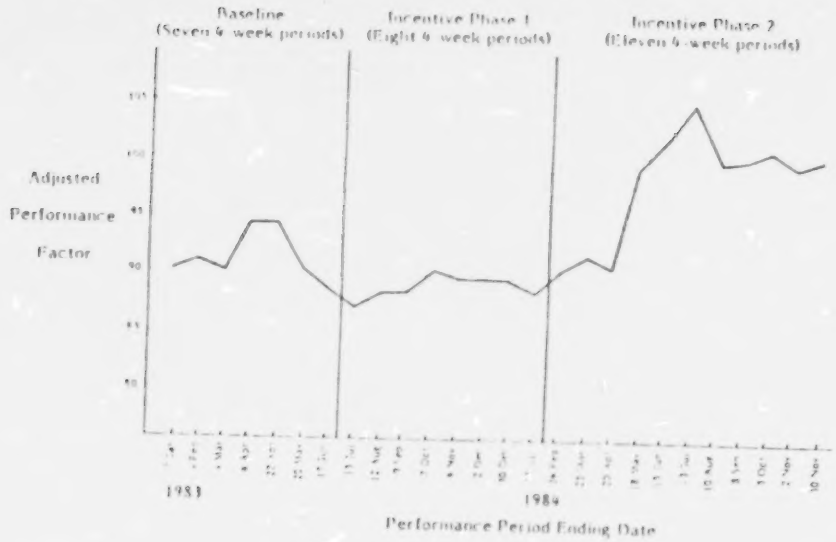


Table 1

Performance and Workload Trends For Three Key Production Shops During Baseline and Two Incentive Phases

	Performance Factor (PF)				
	Baseline[a]	Incentive Phase 1[b]		Incentive Phase 2[c]	
Shop	Average PF[d]	Average PF	% Change From Baseline	Average PF	% Change From Baseline
31	.867	.853	-1.6	.926	+6.8
38	.696	.682	-2.0	.716	+2.9
56	.780	.774	-.8	.792	+1.6

	Workload: Average Man-day Allowances per 4-week Period				
	Baseline	Incentive Phase 1		Incentive Phase 2	
Shop	Average Man-days	Average Man-days	% Change From Baseline	Average Man-days	% Change From Baseline
31	8281	5559	-32.9	5952	-28.1
38	7877	7250	- 8.0	7408	- 6.0
56	7913	7816	- 1.2	7935	+ .3

[a]Baseline: 10 January 1983 - 14 July 1983.

[b]Incentive Phase 1: 15 July 1983 - 27 January 1984.

[c]Incentive Phase 2: 28 January 1984 - 30 November 1984.

[d]Figures represent the average PF within each time frame.

GOAL SETTING WITH NAVY PRODUCTION WORKERS: PROBLEMS AND POTENTIAL

Kent S. Crawford

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Paper presented at the Military Testing Association
San Diego, California, October 1983

ABSTRACT

A goal setting program was implemented in an Navy industrial organization that used engineered performance standards. Results indicated that workers whose baseline performance was below standard set more difficult goals and improved their performance more than high performing workers. Discussion centered on the role of context in influencing goal setting effectiveness in Navy organizations.

BACKGROUND

There is growing concern in the United States with what has been labeled the "U.S. Productivity Crisis" (Newsweek, 1980). This crisis is manifested in the declining rate of growth in the output per hour of labor. The United States finished well behind six other industrial nations in productivity increases from 1968 to 1978 (Bureau of Labor Statistics, 1979). Within the Navy, concern over worker productivity has created increasing interest in productivity improvement at all levels of the organization.

Traditionally, productivity programs in both the military and civilian sectors have centered on technological improvements and capital investments. While the importance of these hardware-oriented approaches is obvious, there is a growing body of organizational literature that suggests that significant productivity improvements can be realized through improved worker motivation (Greiner, Hatry, Koss, Millar, & Woodward, 1981). Several different techniques have been investigated, including autonomous work groups, job restructuring, participative management, and monetary incentive systems. Each of the above approaches has been shown to have merit under differing circumstances (Cummings & Molloy, 1977; Patten, 1977).

Goal Setting

Goal setting is an area of organizational research that seems to be especially promising in terms of enhancing worker motivation and performance. Research has shown that goals are a major source of work motivation (Mitchell, 1979). Likewise, a recent review of field studies using goal setting techniques found a 16% median improvement in worker performance (Locke, Feren, McCaleb, Shaw, & Denny, 1980). Based on the results of a number of highly successful lab and field studies, goal setting has been called "a simple, straightforward, and highly effective technique for motivating employee performance" (Latham & Locke, 1979, p. 80).

While it is clear that goal setting can be an effective motivational technique, we feel that comparatively little study has been directed toward careful investigation of the possible limitations of the approach. There are very likely no panaceas in any field of applied science (Locke, Sirota, & Wolfson, 1976). As such, it would seem that goal setting theory is subject to boundary limitations regarding to whom it applies and where it works best (Miner, 1980). The current study addressed this issue by examining the effectiveness of goal setting in an industrial organization that made extensive use of engineered performance standards.

The use of task standards is an outgrowth of the basic tenets of scientific management (see Taylor, 1967). These standards represent the time a trained employee working at a normal pace would be expected to complete a given task. They are usually based on time and motions studies or on historical performance trends. Within many industrial organizations work standards have been established for most production jobs. While these standards are used for advance cost estimates, manpower projections, and other planning requirements, they also serve another implicit function—they establish acceptable performance levels for workers (Maynard, 1971). In this sense, a standard is a goal for workers to try to achieve (Locke, 1978).

If achieving standards represents an acceptable performance level, then industrial organizations that make extensive use of task standards may encounter problems in implementing goal setting programs for workers. The basic proposition of goal setting theory states that there is a positive relationship between the difficulty of an accepted task goal and level of performance on the task (Locke, 1968). Considerable research has shown that hard, specific goals (if accepted) result in performance improvements (Locke, Shaw, Saari, & Latham, 1981). Performance standards certainly define specific goals; however, they may not always be difficult. While performing at standard level may be challenging for employees with low ability and work motivation, it wouldn't represent a challenging goal for a motivated and highly skilled employee.

Goals Versus Current Performance Levels

The objective of goal setting is to establish specific, challenging goals for all workers. Individuals are encouraged or required to have different goals dependent on their current performance level. The problem with goal setting in an organization using industrial standards is that the organization is sending mixed messages. The supervisor is trying to establish a challenging goal for the worker (often above standard performance level) while the organization has previously defined standard performance as acceptable.

One means of possibly reducing the above problem is for the supervisor to assign goals. The supervisor could then set goals based on current performance independent of existing standards. Research has shown that if goal difficulty is held constant, equal goal acceptance and performance improvements are obtained regardless of whether goals are assigned or set participatively (Dossett, Latham, & Mitchell, 1979; Latham & Saari, 1979; Latham, Steele, & Saari, 1981). However, there is some evidence to suggest that when both participative and assigned goals are set independently, participative goal setting may result in more difficult goals (Latham & Yukl, 1975a; Latham, Mitchell, & Dossett, 1978). However, given the current state of knowledge, it would be difficult to predict which method would be more effective in organizations with existing performance standards. Also, regardless of the method used, it is not certain whether or not workers would set or accept goals above standard.

While the best means of setting goals remains unclear, there is one sub-group of workers who might be expected to improve more as the result of a goal setting program in an industrial organization — low performers. Individuals who are currently

BEST COPY AVAILABLE

performing below standard are not faced with conflicting messages when goals are established by or with the supervisor. In addition, it is possible that high performers are more likely to understand task requirements and have personal performance goals than are low performers. Thus, it seems reasonable to expect a goal setting intervention with production workers to have its greatest impact on low performers.

One recent study supports this contention for nonproduction workers. Pritchard, Bigby, Beiting, Coverdale, and Morgan (1981) found that for data transcribers, goal setting and feedback had a positive impact on poor performers but no impact on good performers. They argued that since the treatment was designed to increase motivation, and since the good performers were probably already motivated, the treatment had little impact on them.

The purpose of the current study was to examine the impact of goal setting and feedback on the performance of Navy industrial employees working within the context of existing performance standards. It was hypothesized that low performers (workers historically performing below standard) would have more difficult goals and would show greater performance improvements than high performers (workers historically performing at or above standard).

METHOD

The engine division of a Naval Air Rework Facility (NARF) served as the research setting for the study. Production workers in this division were involved in the overhaul of aircraft engines, components, and accessories. They were all civil service employees, predominately male, and most had a high school education.

Prior to the goal setting intervention, assigned tasks included a description of required work and the time allocated for the work (i.e., the performance standard). However, although workers knew how well they performed on individual tasks when they completed them, they were not provided with any summary feedback of their performance on all work during a given time period. Since feedback has been shown to be a necessary condition for goal setting to be effective (see Locke et al., 1981), it was first necessary to design an individual work measurement and feedback system.

A computerized system was developed to measure individual level performance using the existing NARF management information system. A weekly report was then generated for each employee providing performance feedback for the previous week. The performance measure was based on how well workers performed against standard and was calculated by taking the ratio of time expended on tasks in a given week to the total standards earned in that same week. This figure was then multiplied by 100. Thus, a rating of 100 meant that an individual completed all work within the standard time allocated. Ratings higher than 100 indicated performance better than standard and those lower than 100, performance below standard.

Twenty-two production shops in the engine division were included in the study. Each shop was supervised by its own foreman. Eleven shops were selected for the goal setting treatment and the remaining 11 were used as a comparison group. All shops were both spatially and structurally distinct subunits.

The 11 experimental shop foremen were trained in the use of the new worker feedback reports and in goal setting; six were trained to assign goals to subordinates while the remaining five were trained to set goals participatively with subordinates. The foremen were asked to arrive at different goals for different subordinates based on the worker's ability, motivational level and current performance. The foremen were initially resistant to the notion of setting challenging goals for workers who were already performing at or above standard. They felt that these employees were currently doing more than should be expected of them. However, the foremen agreed to proceed and give the program a fair chance.

The 11 foremen in the experimental groups met individually with their subordinates to either assign a challenging performance goal or to arrive at such a goal participatively. In addition to receiving the weekly performance report, workers in the goal setting shops met with their foremen individually every 2 to 4 weeks to discuss progress towards their goals and possible work problems.

An 18-week period prior to the beginning of goal setting and feedback was used to establish a baseline level of performance for both the experimental and comparison workers. The 22-week period after program implementation was used to assess program effectiveness. Sixty-seven workers participated in setting their goals while 57 were assigned goals. The comparison shops were composed of 117 workers.

The weekly employee performance data were aggregated to form single pre- and post-treatment performance scores for each worker. Reliability coefficients, computed on the weekly performance measures for the baseline period, indicated that the data were sufficiently reliable for use as overall performance measures (coefficient Alpha = .79).

Research has shown that objective measures of goal difficulty are often better predictors of performance improvement than subjective measures (Yukl & Latham, 1978). For this reason, goal difficulty was operationalized as the difference between an individual's baseline performance score and his/her goal. This measure of goal difficulty allowed for the partial control of baseline individual differences in ability and motivation.

RESULTS

General Results

The initial analyses examined all workers independent of their baseline performance.

Manipulation check. In order to verify the treatment conditions, workers in both the assigned and participative groups were asked to respond on a 4-point Likert scale how much influence they had in setting their goals (1 = a lot of say; 4 = no say). Individuals in the participative condition ($\bar{X}$ = 1.3 reported significantly more influence ($p < .01$) than did workers in the assigned condition ($\bar{X}$ = 3.0)

Performance Change. The mean performance levels for workers in the treatment and comparison groups are presented in Table 1. A test for homogeneity of regression coefficients yielded no significant differences across the groups. Therefore, an analysis of covariance was used to contrast the experimental and comparison groups with the baseline performance measure as a covariate. A significant main effect was found ($p < .01$) indicating differences in treatment performance levels across the groups. Follow-up tests indicated that both the assigned goal setting group (adjusted $\bar{X}$ = 108.3) and the participative group (adjusted $\bar{X}$ = 106.2) were significantly higher ($p < .05$) than the comparison group (adjusted $\bar{X}$ = 101.7). There were no significant differences between the two goal setting groups during the baseline or treatment periods.

BEST COPY AVAILABLE

Table 1

Mean and Adjusted Performance Efficiency Scores

Group	Mean Performance Efficiency: Baseline (B) Period	Test (T) Period	Performance Change (T - B)	Adjusted[a] Test Period	N
Comparison	99.5	102.4	+2.9	101.7 [b,c]	117
Experimental	97.5	106.4	+8.9	107.1 [b]	124
Assigned goals	99.3	108.9	+9.6	108.3 [c]	57
Participative goals	96.0	104.4	+8.4	106.2 [c]	67

[a]Adjusted to control for differences in baseline period performance.

[b]For the analysis contrasting the comparison and combined experimental groups, covariance $F = 11.7$, $p < .001$.

[c]For the analysis contrasting the comparison, assigned goals, and participative goals groups, covariance $F = 6.2$, $p < .01$.

Goal difficulty. A significant correlation was found between objective goal difficulty and degree of performance improvement ($r = .40$, $p < .001$). No difference was found between the average level of goal difficulty in the participative group ($\bar{X} = 8.1$) and the assigned group ($\bar{X} = 11.0$). This finding is consistent with the earlier finding indicating no difference in performance between the two groups in the treatment period. One interesting finding did emerge as to the actual goals set in the two groups. Seventy-three percent of the participative workers had a goal of 100 (or standard level of performance) whereas only 8% of the assigned workers had a goal of exactly 100. This distribution of goals at or different than 100 across the two groups was statistically significant (Chi Square = 30.2, $p < .001$). It thus appears that workers who had some influence in their choice of goals preferred a goal equal to existing organizational standards.

High Versus Low Performers

Workers in the experimental and comparison groups were divided into two categories based on their level of performance during the baseline period: (1) high performers were individuals whose average performance during the 13-week baseline period was at or above standard (i.e., 100), and (2) low performers were individuals whose average performance was below standard.

Performance change. The mean performance levels for high and low performers by different treatment groups are presented in Table 2. Two repeated measures analyses of variance were performed--one for experimental high performers and one for experimental low performers. No main or interaction effects were found for high performers indicating that there was no performance improvement in either the assigned or participative conditions. On the other hand, a main effect for time period was found for the low performers ($p < .01$), indicating that low performers in both the assigned and participative conditions significantly improved their performance as a result of goal setting.

Table 2

Mean Performance Efficiency Scores for High and Low Performers

Group	Baseline Period	Test Period	Performance Change[a]	Adjusted Test Period	N
High Performers					
Experimental	114.0	118.3	+4.3	117.1 [b]	57
Comparison	111.2	111.1	-.1	112.5 [b]	52
Low Performers					
Experimental	83.3	96.3	+12.8	98.3 [c]	67
Comparison	90.1	93.4	+3.3	93.5 [c]	65

[a]The difference between the test period and baseline period performance scores.

[b]Covariance $F = 2.8$.

[c]Covariance $F = 6.3$, $p < .01$.

Because regression towards the mean presented a potential confounding interpretation for the improvements with the low performers, analyses of covariance were also performed on these data. A test for homogeneity of regression coefficients revealed no significant difference among high and low performers across the three groups. Thus, an analysis of covariance was conducted separately for high and low performers using the baseline performance measure as a covariate. The results were identical to those reported earlier. High performers in the experimental and comparison groups did not differ while the low

BEST COPY AVAILABLE

performers in the experimental groups (adjusted $\bar{X}$ = 98.3) were significantly higher ($p < .05$) than the low performers in the comparison group (adjusted $\bar{X}$ = 93.5). Overall, the results suggest than goal setting had a positive impact on the performance of low performers and no effect on high performers.

Goal difficulty. One factor that could explain the different effects of goal setting on low and high performers is goal difficulty. It was proposed that low performers would set (or be assigned) more difficult goals relative to their baseline performance level than would high performers. The results relevant to this hypothesis are given in Table 3. The mean goal difficulty level for all the low performers (15.8) was significantly greater ($p < .001$) than the mean level of goal difficulty for high performers (-.4). On the average, high performers had goals that were slightly lower than their baseline performance level, whereas low performers had average goals that were approximately 16 points above their baseline performance level.

Table 3

Mean Goal Difficulty, Goal Acceptance, and Performance Change for Experimental High and Low Performers

Experimental Group	Mean Goal Difficulty	Mean Performance Change	N
High Performers			
Assigned	5.0	+5.7	27
Participative	-5.3	+3.1	30
Total	-.4	+4.3	57
Low Performers			
Assigned	16.7	+13.0	30
Participative	15.0	+12.6	37
Total	15.8	+12.8	67

Analyses were also undertaken to compare goal difficulty for assigned and participative workers. Results indicated that mean goal difficulty was significantly higher ($t = 2.54$, $p < .05$) for high performers who were assigned goals (5.0) than for high performers who participatively set goals. Indeed, high performers who participatively set goals had an average goal that was more than five points below their baseline performance. No significant difference was found between the mean goal difficulty level of poor performers in the assigned (16.7) and participative (15.0) conditions.

Because the results indicated that a number of workers had negative goals (e.g., goals that were lower than their baseline performance), additional analyses were performed to assess the relation of positive and negative goals to performance change. Results indicated that 31% of the high performers ($N = 28$) had goals that were lower than their baseline performance whereas only 5% of the low performers ($N = 3$) had negative goals. Seventy-five percent of these high performers with negative goals ($N = 21$) were in the participative condition. A significant positive correlation was found between goal difficulty and performance change for low performers with positive goals ($r = .41$, $p < .001$). However, no significant relationships were found for high performers with either negative or positive goals, although the correlation for the later group was marginally significant ($r = .27$, $p < .10$). Also, the performance change scores for high performers with positive goal ($\bar{X} = 8.1$) were higher than those for high performers with negative goals ($\bar{X} = .5$), although this difference was only marginally significant ($p < .10$). These findings suggest that goal setting was somewhat successful for high performers, but only if they had goals higher than their baseline performance.

DISCUSSION

These findings provide support for Locke's (1968) goal setting theory, although they also suggest that goal setting effectiveness may be contingent on contextual factors. First, there was a positive relation between goal difficulty and performance improvement; however, this relationship only held for workers whose goals were higher than their baseline performance. In addition, consistent with the hypothesis concerning low performers and an earlier study by Pritchard et al. (1981), goal setting was more effective with low performances than with high performers. This differential impact was reflected both in terms of greater goal difficulty and more performance improvement.

The goal setting process appeared to be affected by the NARF's use of engineered performance standards. This is supported by the lower mean goal difficulty level for high performers (relative to low performers) that occurred both when goals were assigned and participatively set. Assigned goal setting did result in more difficult goals for high performers than did participative goal setting; however, this difference was not reflected in significant differences in the degree of performance improvement for the two groups. The large proportion of goals that were participatively set at 100 (standard performance level) also suggests that organizational task standards can influence the goal setting process. Workers may have felt that 100% was the most reasonable goal for the organization to expect them to achieve--independent of their baseline performance. With the exception of these findings, participative and assigned goal settings yielded virtually identical results. This is consistent with a large number of lab and field studies (see Locke et al., 1981). The failure to find more difficult goals set in the participatively treatment groups may partially reflect the role of context. Where workers had some influence over their goals, they often opted for what they considered to be fair (i.e., standard) rather than what they felt would be challenging.

Some caveats seem in order. First, the sample size was not large, especially when it was broken down into subgroups. Second, the characteristics of the work force may have played an important role. Navy production workers have more job security than most private sector industrial employees. Thus, these workers may have felt more latitude in choosing negative goals. Finally, goal setting effectiveness was only assessed over 5-1/2 months. There is some evidence to suggest that goal setting effects are not sustained over longer time periods (see Ivancevich, 1976).

BEST COPY AVAILABLE

Overall, the results of this study support the general contention of this paper. Goal setting is an effective motivational technique for Navy production workers but is subject to contextual constraints. In this sense, it has both potential utility and potential problems. There are limitations as to conditions where goal setting works best and for whom it works best (Ivancevich, 1978). There is a need to follow Latham and Yukl's (1975b) suggestion that future research in goal setting begin developing more of a contingency framework. This study was one step in that direction.

REFERENCES

Bureau of Labor Statistics. United States Department of Labor, 1979.

Cummings, T. G., & Molloy, E. S. Improving productivity and the quality of work life. New York: Praeger, 1977.

Dossett, D. L., Latham, G. P., & Mitchell, T. R. The effects of assigned versus participatively set goals, KR, and individual differences when goal difficulty is held constant. Journal of Applied Psychology, 1979, 64, 291-298.

Ivancevich, J. M. Effects of goal setting on performance and job satisfaction. Journal of Applied Psychology, 1976, 61, 605-612.

Ivancevich, J. M. Individual differences and organizational goal setting: Issues, research, and new directions. Paper presented at The National Academy of Management, San Francisco, 1978.

Latham, G. P., & Locke, E. A. Goal setting: A motivational technique that works. Organizational Dynamics, 1979, 8, (2), 68-80.

Latham, G. P., Mitchell, T. R., & Dossett, D. L. Importance of participative goal setting and anticipated rewards on goal difficulty and job performance. Journal of Applied Psychology, 1978, 63, 163-171.

Latham, G. P., & Saari, L. M. The effects of holding goal difficulty constant on assigned and participatively set goals. Academy of Management Journal, 1979, 22, 163-168.

Latham, G. P., Steele, T. P., & Saari, L. M. The effects of participation and goal difficulty on performance. TR-GS-6. College Park, Maryland: University of Maryland, 1981.

Latham, G. P., & Yukl, G. A. Assigned versus participative goal setting with educated and uneducated woods workers. Journal of Applied Psychology, 1975a, 60, 299-302.

Latham, G. P., & Yukl, G. A. A review of research on the application of goal setting in organizations. Academy of Management Journal, 1975b, 18, 824-845.

Locke, E. A. Toward a theory of task motivation and incentives. Organizational Behavior and Human Performance, 1968, 3, 157-189.

Locke, E. A. The ubiquity of the technique of goal setting in theories and approaches to employee motivation. Academy of Management Review, 1978, 3, 594-601.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N., & Denny, A. T. The relative effectiveness of four methods of motivating employee performance. In K. Duncan, M. Gruneberg, & D. Wallis (Eds.), Changes in working life. New York: Wiley, 1980.

Locke, E. A., Sirota, D., & Wolfson, A. D. An experimental case study of the successes and failures of job enrichment in a government agency. Journal of Applied Psychology, 1976, 61, (6), 701-711.

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. Goal setting and task performance: 1969-1980. Psychological Bulletin, 1981, 90, (1), 125-152.

Maynard, H. B. (Ed.). Industrial engineering handbook. New York: McGraw, 1971.

Miner, J. B. Theories of organizational behavior. Hinsdale, IL: Dryden Press, 1980.

Mitchell, T. R. Organizational behavior. Annual Review of Psychology, 1979, 30, 243-281.

Newsweek. An economic dream in peril. Author. September 8, 1980, 50-69.

Patten, T. H. Pay: Employee compensation and incentive plans. New York: Free Press, 1977.

Pritchard, R. D., Bigby, D. G., Beiting, M., Coverdale, S., & Morgan, C. Enhancing productivity through feedback and goal setting. AFHRL-TR-81-7. Brooks AFB, TX: Air Force Human Resource Laboratory, May 1978.

Taylor, F. W. The principles of scientific management. New York: Norton, 1967. (Originally published, 1911).

Yukl, G. A., & Latham, G. P. Interrelationships among employee participation, individual differences, goal difficulty, goal acceptance, goal instrumentality and performance. Personnel Psychology, 1978, 31, 315-323.

The views expressed in this paper are those of the author and are not necessarily those of the Department of the Navy.

BEST COPY AVAILABLE

THE DETERMINANTS OF GOAL CHOICE, WORK MOTIVATION AND TASK PERFORMANCE

James A. Riedel
Delbert M. Nebeker
Navy Personnel Research and Development Center
San Diego, California 92152-6800

Background

In many organizations goal setting has been found to be a powerful technique for influencing work motivation and performance (Locke, Feren, McCaleb, Shaw, and Denny, 1980). While the positive effects of goal setting on worker productivity have been demonstrated repeatedly, a major weakness of this approach is the failure to specify the process by which goals are set. The central argument of this paper is that the process of goal choice may be central to understanding the relationship among organizational context, goal setting, motivation and performance.

With few exceptions, there has been little research directed toward understanding the determinants of goal choice, acceptance, and commitment (Steers & Porter, 1979). This is unfortunate because goal setting does not take place independently of the work place; reward systems and other work setting characteristics come together to affect goal choice, acceptance, and commitment (Crawford, 1982). Some investigators, however, have attempted to use an expectancy theory model to explain goal choice and acceptance. For example, in both laboratory and field settings goal acceptance has been reliably predicted using expectancy and valence measures (Dachler & Mobley, 1973; Mento, Cartledge, & Locke 1980; Steers, 1975). In a related line of research, expectations of success and the value placed on the outcomes of goal attainment were found to be the principal determinants of "level of aspiration" (Frank, 1941; Hilgard, 1942/1958). These two factors are highly related to the core concepts in expectancy theory: expectancy and valance (Vroom, 1964).

In addition to a limited understanding of how people set goals, the relationship between goal setting and other motivational techniques is unclear. This, in part, may be due to the fact that there has been little integration of goal setting with motivational theories. The role of motivational techniques, such as monetary incentives and goal setting, in work motivation and performance is one of the most underresearched and poorly understood areas in organizational behavior (Opsahl & Dunnette, 1966; Lawler, 1981).

The need for a better understanding of the process of goal choice is evident. This process may provide insight into the relationship among goal setting, organizational context, motivation, and performance. Also, this process, if linked to motivation theory, should help to clarify the relationship among goal setting and other motivational techniques, such as monetary incentives. The purpose of this study is to offer a preliminary model of goal choice, work motivation and performance. This model is presented in Figure 1. As a preliminary test of its validity, selected elements in the model will be examined to determine its usefulness as an explanatory device.

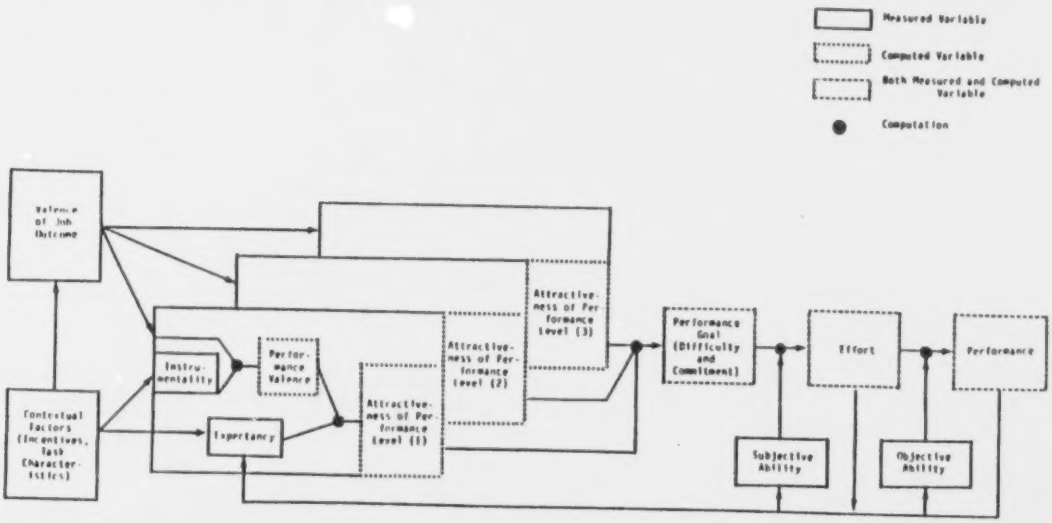


Figure 1. A Model of Goal Choice, Work Motivation, and Performance.

Work Motivation Model

Based on expectancy theory concepts and processes, the model shown in Figure 1 explains work motivation and performance as a cognitive process where an individual chooses, from alternative performance goal levels, the level perceived to be most attractive. This perception of attractiveness is based on various beliefs and feelings a person has regarding the likelihood that performing at certain levels will lead to particular job outcomes. Contextual factors, such as the opportunity to earn monetary incentives for good performance, will influence these beliefs and feelings. The hypothesized effect of the performance goal is to influence the amount of effort a person is wiling to expend in accomplishing the goal. Furthermore, an individual's self-assessment of ability as well as actual ability are presumed to moderate the relationships among performance goal, effort, and performance. Since in this model the goal concept is a major determinant of effort and performance, it is crucial to understand how people choose their performance goals.

BEST COPY AVAILABLE

The model suggests that contextual factors influence valence, an individual's anticipated satisfaction with particular levels of different job outcomes, and instrumentality, the expectancy that different performance levels are associated with different outcomes. Contextual factors also affect expectancy, a person's belief concerning the likelihood of achieving a particular level of performance if they tried their best. Valence and instrumentality combine multiplicatively to determine performance valence. Performance valence is a hypothetical construct that represents the anticipated satisfaction of performing at a given level of performance. The anticipated satisfaction for a given performance level is derived from its degree of association with particular job outcomes and the valence of those job outcomes to the individual. Performance valence combines multiplicatively with expectancy. This product becomes the perception of attractiveness for each performance level. In essence, each possible level of task performance acquires valence through its association with certain job outcomes and the anticipated satisfaction associated with these outcomes. This performance valence is then modified by a person's belief concerning the likelihood of achieving that level of performance given his or her best effort. The result is a perception of attractiveness for each level of performance.

The model specifies that goal choice is based on a person's evaluation of the relative attractiveness of various performance levels. The model is flexible in that it accommodates alternative decision strategies (e.g., return on effort, maximization, value matching). In this study the return on effort approach, which assumes that people use an incremental decision rule in choosing a goal, was used in determining the goal choice prediction. With this approach the model would predict goal choice to be some measure reflecting the marginal gain in the attractiveness of performance for performing at a particular level. While this approach has been successfully employed to improve expectancy theory predictions of performance (Kopelman, 1977), empirical evidence is lacking concerning the relative accuracy of alternative decision strategies. Additional work is needed to determine whether the return on effort approach offers the best representation of the goal choice process.

The remainder of the model describes the process by which performance goals are translated into work motivation, and the work motivation into task performance. The hypothetical relationship of these concepts and the process by which goals are translated into performance is based on the broad theoretical position that Performance (P) equals the product of Ability (A) and Motivation (M); (P = A X M).

Though cognitions serve telic purposes, they are influenced by past behavior and experience. The model specifies feedback loops suggesting that a person's effort-performance expectancy is influenced by past expenditures of effort and performance. Also, past performance affects both a person's objective ability and their subjective estimate of their ability. These factors, in turn, affect future effort and performance. The model is dynamic in that the source of purposive action is cognitive activity, though not necessarily conscious, that is influenced by past action and its consequences.

Method

Subjects

One-hundred and thirty experimental subjects participated in this study. Their average age was 21 years. Seventy-one of the subjects were female and 59 were male. Approximately 40% were high school students and the remaining 60% were undergraduate college students. Some data for six subjects were missing and therefore were unavailable for some of the analyses.

Procedure

The present study was part of a larger work simulation study designed to investigate the effects of alternative incentive magnitudes on performance (Riedel, Nebeker, & Cooper, 1985).

Subjects were recruited for part-time employment to perform a clerical transfer task. The 130 subjects who qualified for the job were assigned randomly to 1 of 7 experimental conditions differing in terms of the magnitude of incentive offered for various levels of performance. They worked 5 days, 4 hours a day, for a total of 20 hours at a rate of $4.40 per hour.

Research questionnaires were administered three times: after assignment to an experimental condition, at the start of the third day, and at the start of the fifth day. These questionnaires contained the expectancy and goal items needed for evaluating the model predictions. The quality and quantity of performance was recorded daily. A detailed description of the experimental procedure, treatment conditions, constructs and measures, and method of wage and incentive payment can be obtained from the author.

Results

Manipulation Check

Incentives and performance. It was expected that subjects in the incentive conditions would perform better than subjects in the nonincentive groups. To test this hypothesis, an analysis of variance was performed with treatment condition as the dependent variable.

The results of this analysis suggest a significant treatment effect, $F(6,120) = 3.27$, $p < .005$. A planned comparison of the performance means for the incentive and nonincentive groups revealed a significant difference, $t(120) = 3.87$, $p < .001$.

Incentives and instrumentality. Subjects were asked the amount of pay they expected to receive if they were to perform at alternative levels of performance. Judging from the responses, the performance-pay relationship was accurately perceived by most subjects across the treatment groups. For all conditions the reported pay instrumentalities approximate the actual relationships between pay and performance.

Model Predictions

The central research question pertained to the capacity of the model to account for the process of goal choice and task performance. To evaluate the model, goal choice (level) was predicted by the model, using a return on effort decision algorithm. This prediction was compared with self-reported goal choice. Also, the ability of the model to predict performance was evaluated by correlating the predicted performance with actual task performance. The results of these analyses are summarized below, first for goal choice and then task performance.

Goal choice. Prior to evaluating the goal choice prediction, responses to the self-report goal choice question were examined. Twenty-six subjects selected a single quantitative goal (e.g., 5 units per hour), 39 subjects selected a quantitative

113 BEST COPY AVAILABLE

goal with a range (e.g., 5-7 units per hour), and 60 subjects set non-quantitative goals (e.g., to do my best). To increase the sample size for the single quantitative goal category, single goals for subjects with quantitative range goals were computed by averaging the upper and lower anchors of their range goal response. This resulted in 65 cases with a single quantitative production goal.

The model prediction of goal choice correlated significantly with self-reported goal choice (r = .81) and task performance (r = .50), both significant ($p < .001$). It was of interest to determine if the relationship between this goal measure and performance maintained for the entire sample, including people without quantitative goals. For the entire sample the correlation between the predicted goal choice and performance is (r = .54) and for people without quantitative goals (r = .57), both significant ($p < .001$). Also, the results of a t test indicate no significant mean difference in performance between those people who set quantitative goals and those people who did not set a quantitative goal. It appears that the goal choice prediction from the research model, based on the return on effort algorithm, relates highly to self-reported goal choice. The model also provides a significant prediction of performance, regardless of whether a subject reported setting a quantitative goal.

Performance. The capacity of the model to predict performance was evaluated by correlating the predicted performance with actual task performance. The model prediction of performance was significantly correlated with actual performance on the task (r = .46, $p < .001$). While this prediction was slightly better for subjects who set quantitative goals (r = .55, $p < .001$) than subjects who did not (r = .39, $p < .001$), the difference between these correlations was not significant. These findings provide preliminary support for the validity of the research model in predicting performance.

Discussion

The central purpose of this study was to improve our understanding of the process of goal choice. Overall the goal choice process specified in the research model was supported by the findings. First, monetary incentives were found to influence pay instrumentality. Second the cognitive components of goal choice which were specified in the research model predicted self-reported goals and performance, suggesting that the process of goal choice may be linked to expectancy theory concepts and processes.

The effect of incentives on pay instrumentality indicates that the treatment affected individual perceptions about the amount of pay associated with alternative levels of performance. Results show that the pay instrumentalities approximate the actual relationships between pay and performance, indicating that the pay contingencies were perceived quite accurately. In terms of the model, the effect of the treatment was to increase instrumentality and thereby increase performance valence, the anticipated satisfaction of performing at a given level of performance. The performance valence for a given performance level is derived from its degree of association with particular job outcomes and the valence of those job outcomes to the individual. It can be concluded that the experimental treatment was very effective in strengthening this association.

The goal choice process specified in the research model was supported by the findings, suggesting that expectancy theory concepts may be useful in understanding the cognitive components of goal choice. The combination of the expectancy constructs produced a reasonably accurate prediction of goal choice. The predicted goal choice was significantly correlated with the actual self-reported goal. This finding suggests the interpretation that goal choice is a cognitive process where an individual chooses, from alternative performance goal levels, the level perceived to be most attractive. This perception of attractiveness is based on various beliefs and feelings a person has regarding the likelihood that performing at certain levels will lead to particular job outcomes. The results indicate that contextual factors, in this case the opportunity to earn monetary incentives for good performance, influence these beliefs and feelings.

This study has contributed to a better understanding of the relationship between organizational context and goal setting as they relate to work motivation and performance. The findings suggest that the process of goal choice is central to understanding how contextual variables influence goals, motivation, and performance. Moreover, the research model provides a useful starting point for investigating the relationships between organizational context and employee cognitions and perhaps for integrating goal setting with expectancy theory.

References

Crawford, K. S. (1982). Goal setting with industrial workers: The impact of contextual factors. Unpublished doctoral dissertation, University of California, Irvine, California.

Dachler, H. P., & Mobley, W. H. (1973). Construct validation of an instrumentality-expectancy-task-goal model of work motivation: Some theoretical boundary conditions. Journal of Applied Psychology, 58, 397-418.

Frank, J. D. (1941). Recent studies of the level of aspiration. Psychological Bulletin, 38, 218-226.

Hilgard, E. R. (1958). Success in relation to level of aspiration. In C. L. Stacy and M. F. DeMartino (Eds.), Understanding human motivation (pp. 235-241). Cleveland: Howard Allen. (Original work published 1942).

Kopelman, R. E. (1977). Concepts, theories and techniques across-individual, within-individual and return on effort versions of expectancy theory. Decision Sciences, 8, 651-662.

Lawler, E. E., III. (1981). Pay and organization development. Reading, MA: Addison-Wesley.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N., & Denny, A. T. (1980). The relative effectiveness of four methods of motivating employee performance. In K. D. Duncan, M. M. Gruneberg, & D. Wallis (Eds.), Changes in working life: Proceedings of the NATO International Conference (pp. 1121-1157). London: Wiley.

Mento, A. J., Cartledge, N. D., & Locke, E. A. (1980). Maryland vs. Michigan vs. Minnesota: Another look at the relationship between expectancy and goal difficulty to task performance. Organizational Behavior and Human Performance, 25, 419-440.

Opsahl, R. L., & Dunnette, M. D. (1966). The role of financial compensation in industrial motivation. Psychological Bulletin, 66, 94-118.

Riedel, J. A., Nebeker, D. M., & Cooper, B. L. (October 1985). The influence of monetary incentives on goal choice, goal commitment and task performance (NPRDC Tech. Rep. In review). San Diego: Navy Personnel Research and Development Center.

BEST COPY AVAILABLE

Steers, R. M. (1975). Task-goal attributes, n achievement, and supervisory performance. Organizational Behavior and Human Performance, 13, 392-403.

Steers, R. M., & Porter, L. W. (1979). Motivation and work behavior. New York: McGraw-Hill.

Vroom, V. H. (1964). Work and motivation. New York: Wiley.

The views expressed in this paper are those of the authors and does not necessarily reflect the views of the Department of the Navy.

BEST COPY AVAILABLE

THE EFFECTS OF REWARD MAGNITUDE AND DIFFICULTY OF PERFORMANCE STANDARDS UPON INDIVIDUAL PRODUCTIVITY

Delbert M. Nebeker[1]

Navy Personnel Research and Development Center
San Diego, CA 92152-6800

Management systems that control productivity and performance are critical to the success of organizations. One form of control system known as a Performance Contingent Reward System (PCRS), is receiving increasing attention in the U.S., particularly when it involves financial incentives. This is partly a function of the critical productivity problem we face in this country, and partly because recent evidence shows financial incentives to have a strong positive impact on performance (e.g. Nebeker, Neuberger, 1985; Locke, Feren, McCaleb, Shaw & Denny, 1980). In spite of this evidence the use of financial incentives as a means to increase productivity remains a controversial issue (Belcher, 1974; Lawler, 1983). If financial incentives are to be used effectively as a means to improve worker efficiency, we must have a better understanding of how they operate. We need to know how reward systems should be designed to maximize their value.

The design of reward systems can vary along a number of different dimensions. Theses include the following:

1. Objectivity of performance measure. The degree to which the performance measure is measured objectively as opposed to subjectively.

2. Performance aggregation level. The number of people include in the performance measure who share a reward.

3. Performance standard. The difficulty of the performance level required to earn a reward.

4. Sharing rate. The percent of earnings "saved" by performing above standard that is given as a reward.

5. Performance period. The length of time that performance data is accumulated before a reward determination is made.

6. Feedback type. The method of providing performance feedback.

7. Feedback period. The length of time between performance feedback.

---

[1]The views expressed in this paper are those of the author and are not official and are not necessarily those of the Department of the Navy

8. Performance-reward function. The shape of the function relating reward to performance.

9. Incentive period. The length of time following the performance period before payment is made.

Little is known about the optimal values for these parameters in various work situations. Virtually no empirical research directly addresses these issues in work environments. Theory is only slightly more helpful. If we were going to get answers about the optimal values for these parameters we were going to have conduct some parametric studies. Our field experience had suggested that two of the more important parameters in the design of reward systems were standard difficulty and reward magnitude.

In reviewing the literature we found the following: There is disagreement over whether performance standards ought to be made difficult to reach (Locke, et al., 1981, Barnes, 1980) or attainable for most workers and therefore easy for many (Peters & Waterman, 1982, Motowidlo, et al., 1978). Evidence for both points of view can be cited. One possible reason for the apparent contradiction is that the effects of the reward magnitude have not been adequately considered in research on standards or goal difficulty. It is quite likely then, that the affects of easy or difficult performance standards are moderated by the magnitude and/or the attractiveness of the rewards available for reaching and exceeding these standards (Matsui, Okada & Mizuguchi, 1981).

The instrumental learning and conditioning literature (c.f. Logan, 1970,p.90-91) posits that increasing magnitudes of reward have "diminishing returns" on performance (at least for rats). This suggests that very large rewards are likely to have a lower marginal utility than moderate rewards. Some of our own preliminary research supports this contention with people at work. Practical applications of reward systems in real organizations, however, show wide variability in the amounts of reward offered for performance above standard. Examples in business and industry can be found with sharing rates ranging from 10% to over 100%. It is reasonable to assume then that systems designed to pay very large rewards are likely to produce marginally less improvement and be less cost effective than systems that pay more moderate amounts.

The present research was designed to help us understand the interactive effects of these two variables by exploring the joint effects on productivity of varying degrees of reward magnitude, as defined by sharing rate, and standard difficulty. Furthermore, it is expected that worker ability will affect the relationship between standard difficulty and reward magnitude.

Method

Research setting

This research was conducted in the Organizational Systems Simulation Lab (OSSLAB) at the Navy Personnel Research and Development Center. The OSSLAB is designed to create a high fidelity simulation of real computerized work environments. Microcomputers are used as workstations so that individuals can be hired to do real work tasks under experimentally controlled conditions.

Subjects and design

Twenty-four employees (8 males,16 females) were recruited and hired (at $4.89 per hour) to provide technical support to the Navy Personnel Research and Development Center. Their job was to enter and maintain references in a data base for searching and retrieving the scientific literature. The Ss were required to be keyboard proficient before being hired. They worked two 4-hour shifts a week for eight weeks. A Work sample test given to measure ability revealed no significance differences between the two shifts.

The research design called for the Ss to perform their work under three reward conditions: (1) Baseline or control; (2) small incentives, wherein 15% of the wages saved by performing above standard were paid to the employee as a bonus and; (3) large incentives, wherein 50% of the wages saved by performing above standard were paid to the employee as a bonus. Furthermore, the shifts were designated as either the easy standard group or the difficult standard group. The easy standard group had a performance goal or standard set at the 20th percentile of the group's baseline performance level. This meant that 80 percent of the group was already exceeding the standard when the incentives were introduced. The difficult standard group had their performance goal, or standard, set at the 90th percentile. This meant that only 10 percent of the group were exceeding the standard when the first incentive was introduced. These values were chosen to match two interesting findings. Barnes (1980) demonstrates that standards set by usual industrial engineering methods typically produce standards that only 10% of the people exceed without performance standards and feedback or incentives. The difficult standard shift was designed to match this condition. The easy standard was chosen to match the finding that 80% of workers believe they are performing above average. Thus the standard would be consistent with there own self-concept.

The final design was a 2 X 3 factorial design where standard difficulty level was a between-subject factor and incentive level was a within-subject factor.

Procedure

On the first day of work the Ss reported to the job site and were welcomed by their first and second level supervisors. They

were given an overview of the work they would be doing with an emphasis being placed upon the value of their job and how important accuracy and quality were to the information system a long with speed.

They were then trained on the use of their work stations, (IBM PC-XT microcomputers). This included the use of both hardware and software. Upon completion of this training their shift was completed for the day and they were excused to leave. On the second day they were given additional training on the task, allowed several practice items and then given the work sample test. Following the work sample they began their assigned work of entering and maintaining the references in the data base. The remainder of their employment consisted of their performing the task each workday. The only variations in this schedule were: (1) the introduction of the small incentive during the third week and the introduction of the large incentive during the sixth week; and (2) the administration of work perception questionnaires on four occasions spread throughout their eight weeks of employment; (3) the readministration of the work sample test at the end of the second week.

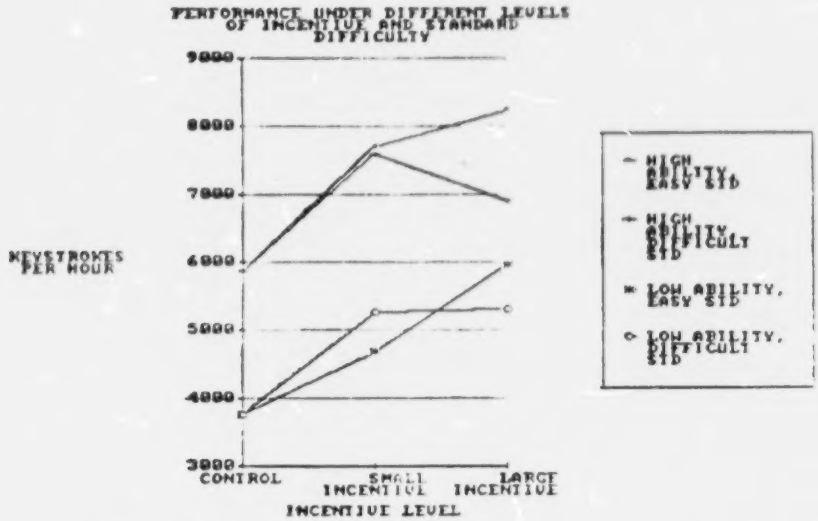
Performance was measured by keystroke rate, the number of keystrokes per hour. At any time during the experiment the workers could chose to view, on their screens, one of several reports of their current and "to-date" performance. During the baseline/control condition, prior to the introduction of the incentives, these reports included only raw performance information such as keystroke rate, hours on the tasks and regular pay. Following the introduction of the incentives, however, the reports added: (1) A listing of standards and the workers current and to-date performance efficiency against these standards (e.g. keystrokes per hour/standard keystrokes per hour); (2) the current and to-date bonus earned for exceeding the performance standards; (3) current and to-date total earnings

Results and Discussion

Daily performance means were used as the chief dependent variable in a series of moderated multiple-regression equations. In these equations the ability score was entered first, as a subject-covariate factor; then the the dummy coded treatment main effects; followed by the two-way interactions and; finally the three-way interactions.

The results of these analyses reveal a highly significant multiple R at each step and an overall R=.89 (R^2=.79; p<.001). In this regression three significant predictors of keystroke rate were found. Ability, as measured by the work sample test, accounted for a large portion of the performance variance (Beta=.583;F=42.42;df=11/202; p<.001). Also contributing significant portions of the variance to the prediction of performance were the interaction of standard difficulty with the large incentive manipulation, (Beta=1.213;F=8.21;df=11/202;

p<.005) and a three-way interaction between ability, standard difficulty, and large incentives (Beta=-1.274;F=9.928;df=11/202; p<.005). The relationships produced from these effects are shown in the figure below.



As can be seen the independent variables and their interactions produce some interesting effects on keystroke rate. First, it is obvious that there are large ability differences in performance and interaction effects for incentive level, standard and ability. High ability workers perform significantly better in all conditions than low ability workers. Second, when performance standards are relatively easy, increases in incentives result in performance increases for all ability levels at each level of incentive. When the standards are difficult, however,_ performance for_the_high ability workers, first increases for the small incentives and then <u>decreases</u>, when the large incentives are introduced. Finally, with small incentives and the difficult standards, the low ability workers perform better than their counterparts with easy standards. With the introduction of the large incentives, however, their performance, does not improve, while the easy standard group's does, allowing the easy standard group to out- perform the difficult standard group.

From these results the following points appear clear: (1) Performance improves substantially with the introduction of both standards and incentives regardless of whether or not the standards are difficult or easy and whether or not the incentives are large or small. (2) It appears that the amount of improvement with easy or difficult standards depends upon whether or not small or large incentives are being offered for exceeding them. The best performance, contrary to Locke et al.'s position (1981), occurs when the standards are at the 20th percentile and the sharing rate is 50%. (3) High ability worker's behavior

under conditions of both high standards and large incentives seem to be either discouraged by the opportunity to earn incentives, or discount the value of increasing performance and actually reduce their performance. Either of these explanations requires a view of small incentives as providing only an achievement or competence motive for reaching the goal. When large incentives are introduced, however, financial motives for reaching the goal may become dominant. Under these conditions the question of discouragement and/or relative value may be significant factors affecting performance. Which of these explanations (or some other possibility) is superior remains for future research to determine.

References

Barnes, R. M. (1980). Motion and time study: Design and measurement of work. New York: Wiley.

Belcher, D. W. (1974). Compensation administration. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Lawler, E. E., III. (1983). What ever happened to incentive pay? (Tech. Rep. GB3-6(37)). Los Angeles: University of Southern California, School of Business Administration.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N., & Denny, A. T. (1980). The relative effectiveness of four methods of motivating employee performance. In K. D. Duncan, M. M. Gruneberg, & D. Wallis (Eds.), Changes in working life: Proceedings of the NATO International Conference (pp. 1121-1157). London: Wiley.

Logan, F. A. (1970). Fundamentals of learning and motivation. Dubuque, IA: Wm. C. Brown.

Matsui, T., Okada, A., & Mizuguchi, R. (1981). Expectancy theory prediction of the goal theory postulate, "the harder the goals, the higher the performance." Journal of Applied Psychology, 66, 54-58.

Motowidlo, S. J., Loehr, V., & Dunnette, M. D. (1978). A laboratory study of the effects of goal specificity on the relationship between probability of success and performance. Journal of Applied Psychology, 63, 172-179.

Nebeker, D. M., & Neuberger, B. M. (1985). Productivity improvement in a purchasing division: The impact of a performance contingent reward system. Evaluation and Program Planning, 8,121-134.

Peters, T. J., & Waterman, R. H., Jr. (1982). In search of excellence: Lessons from america's best-run companies. New York: Harper & Rowe.

Results of an External Evaluation System Development in the Canadian Forces

Commander R.H. Kerr CF
and
Mr. Duane Tyerman

Background

The External Evaluation Process as part of the ISD (Instructional Systems Development) model is under active study within several commands in the CF (Canadian Forces). One such study, which is still ongoing, was described in MTA Proceedings (Kerr et al, 1984); this short addendum reports results and analysis of the findings in the studies involving 79 Dental Technicians (graduates and supervisors) and 467 Administrative Clerks (graduates and supervisors). Action proposed and underway is discussed in the last portion of this report.

System Description

The system employed a mail-out questionnaire approach (individuals also responded by mail). Question formats required the graduates and supervisors to assess task completion in accordance with a question grid providing for eight possible response patterns per task. Both trades were assessed on 40-50 tasks. Response sheets were manually input into computer files and subsequently edited and analysed using special programmes designed specifically to be 'user digestible'. Details of the formats and outputs are described in the previous paper.

Results - Input Analysis

Ninety-eight percent of the questionnaires were returned and 42% of the response sheets contained recording errors. Further analysis resulted in only 18% of task responses being spoiled out of a total of 43,680. Although an 18% spoilage rate is considered acceptable by the authors, revision to the input formats have been made and considerable improvement is expected in reducing recording errors. A contract has been awarded to a research firm to utilize optical mark reader equipment incorporating the editing function. Because of the relative uniqueness of the input format, and the multi-track aspects of response patterns possible in such a paper-based system, no other option appears available in a 'mail-out' setting. The majority of personnel responding found the format easy or very easy to complete and spent a mean time of 32 minutes completing the questionnaire.

As indicated in the previous paper, test-retest reliability was attempted with a sample of 60 of the original population, by a readministration of the same instrument one month later. After editing, 74% of responses were identical to the original administration of those sampled. Because of the multi-track aspects of response possibilities and because some response change is expected (e.g. personnel may now have performed a task not performed at the first administration) this response is taken to be reasonably reliable, and with form revision, line correlations should improve.

Processing and Outputs

Processing was described in the previous paper and outputs were generated and provided to Training Managers. The External Evaluation business could provide some earth-shaking insights into training continua but not in these instances! Cases involving performance deficiencies beyond the limited level were virtually non-existent and those tasks reported as being not required were easily explained away as being infrequently performed. A good example was the completion of a casualty form by a junior administrative clerk. Infrequently completed tasks such as these are being examined for the degree of emphasis that these tasks have on course lengths. Further action may be taken to alter job specifications which may eventually effect the level at which the training is conducted. The authors are convinced that their next populations should be chosen to expose and quantify areas where some more serious problems are known to exist, which would enable more components in training continua to be examined.

Training Managers on the whole were pleased with and understood the outputs. The emphasis on job-orientation utilized in this system took much explaining, mainly because the audience for outputs were training managers, not performance technologists as discussed in the previous paper.

Further Analysis and Direction

The question of external evaluation in the CF has received renewed interest in 1985 at National Defence Headquarters and other Commands have instituted their own programmes and studies. The authors believe that External Evaluation as applied to Pilot Training may demand different emphasis than that required for Naval Technician training in that output analyses have different descriptors in order for alleviable action to take place in correcting performance deficiencies. The major demand appears to be in developing a system which will monitor and correct 'overtraining' in order to optimize training system efficiency. Apart from additional questions being asked in the field using an existing External Evaluation system to 'flag' a possible overtraining area, the authors at this stage feel that the internal evaluation and professional design processes would provide more impact in guaging and correcting overtraining. Experimental designs could then be utilized to verify, with the assistance of an External Evaluation System whether differing training methods or standards result in acceptable field performance. The authors have been tasked to examine this area in 1986.

Conclusion

This brief resumé of proceedings is intended to provide interim results on the development of an External Evaluation system within a CF context. Future developments using improved technological advances such as optical scanning techniques and use of video display terminals as inputs to external evaluation systems harbinge a heightened future for this often neglected process.

References:

Kerr, R.H. et al (1984). External Evaluation Revisited - an Experimental Update. Proceedings of the 26th Military Testing Association, 1984, 2, 763-768.

Occupational Learning Difficulty: A Construct Validation Against Training Criteria

Joseph L. Weeks
Air Force Human Resources Laboratory

Michael D. Mumford
Georgia Institute of Technology

Francis D. Harding
Advanced Research Resources Organization

Occupational learning difficulty is defined as the time required to learn to satisfactorily perform occupational tasks. It is expressed in terms of a quantitative index which is produced on the basis of information obtained from structured job analysis questionnaires developed and administered by the United States Air Force Occupational Measurement Center. For any given job specialty, a learning difficulty (LD) index is derived by combining relative ratings of task learning difficulty obtained from senior-level technicians and benchmark ratings of task learning difficulty obtained from external occupational experts. The adjusted task ratings resulting from this combination are then weighted and aggregated to produce an occupational-level index of learning difficulty which can be meaningfully compared across job specialties. A more detailed description of the derivation procedure has been provided elsewhere (Weeks, 1984).

Once an LD index was available for most Air Force enlisted job specialties and its reliability and validity had been demonstrated, it served as a job-centered, frame-of-reference for various management decisions. In this context, there are both primary and special applications of learning difficulty information. Primary applications involve both personnel and training management.

For example, to the extent practical, the order of job aptitude requirement minimums are established so as to correspond to the order of job specialties in terms of learning difficulty. This application contributes to the optimal allocation of talent by ensuring that job specialties which are highest in learning difficulty are manned by enlistees having the highest aptitudes.

Learning difficulty indexes are also applied during the initial job-offer process. Air Force enlistees are assigned to job specialties at military entry processing stations on the basis of a computer-based, person-job match algorithm known as PROMIS (Hendrix, Ward, Pina, & Haney, 1979). One of the policies implemented by PROMIS is to offer the most difficult job specialties to the most talented enlistees. This process obviously requires information concerning enlistee aptitudes and information concerning job difficulty. Within the PROMIS system, job difficulty is defined in terms of the LD index.

Another primary application involves decisions concerning mode of training. Enlisted job specialties are designated as either category A, B, or C skills. For category A skills, the mode of training is formal,

resident or school-house training. For category C skills, the mode of training is on-the-job training (OJT). For category B skills, the mode of training can be either formal resident training or OJT. The LD index is one of several inputs to the decision process associated with determining mode of training.

In addition to these primary applications, there have been noteworthy special applications. The LD index was used as an empirical basis for justifying Air Force personnel quality standards in response to inquiries by the House and Senate Armed Services Committees during the development of the 1985 Defense Appropriations Bill. Furthermore, it has been advanced as an empirical basis for Air Force job and training requirements during investigations by oversight committees such as the Government Accounting Office and the Air Force Audit Agency.

Problem

Because of the importance of such applications, the validity of the LD index is an issue of considerable interest. Burtch, Wissman, and Lipscomb (1980) were the first to seriously address this problem. Their approach consisted of correlating relative ratings of task learning difficulty by senior-level technicians with benchmark ratings by occupational experts. Observed correlations ranged from .54 to .96 for 100 different job specialties. These results provided evidence of the convergent validity of the benchmark ratings. Although this study was comprehensive in that separate analyses were conducted for several different job specialties, it cannot be considered sufficient by itself. Validation efforts must take into account the functional role of learning difficulty information in management applications. For all the applications previously described, the occupational-level index rather than task-level ratings of learning difficulty served as the referent for management decisions. Consequently, there appears to be a need to evaluate the validity of the occupational-level index of learning difficulty.

Method

Because the LD index is not applied to predict some criterion, construct validation is considered to be more appropriate than predictive validation. With construct validation, the goal is to evaluate the intrinsic meaning of some measure of interest. This is accomplished by evaluating both convergent and discriminant validity (Campbell & Fiske, 1959). Because the LD index is a measure of learning time, strong relationships with the training time for an initial-skills course associated with a specialty would be expected. Also, because the LD index is a measure of a job property, strong relationships with measures of personnel attributes would not be expected.

In an independent research project devoted to the development of a covariance-structure model of Air Force technical training, it was necessary to collect information concerning occupational learning difficulty as well as measures of student input, course content, and training outcome variables for several initial-skills courses. Therefore, it was decided to extend that effort to include an examination of the construct validity of the LD index. Only analyses relevant to the construct validation of the LD index

will be discussed here. Details concerning the development of the covariance-structure model of initial-skills training are provided by Mumford, Weeks, Harding, and Fleishman (1985).

Samples of subjects and courses used for the present analyses were identical to those employed by Mumford et al. (1985). Courses were selected from among approximately 200 initial-skills courses administered by various technical training centers under Air Training Command. Courses were selected to provide a representative sample with respect to numerous criteria including course content, student flow, training costs, and aptitude area. Subjects were sampled so as to provide a minimum of 50 students for each course. These procedures provided a total of 5,970 students and 48 initial-skills courses.

The independent variable examined in the present study consisted of the LD index obtained for the job specialty associated with each of the 48 initial-skills courses. For each specialty, the index consisted of an aggregate value obtained by deriving the average learning difficulty of first-term positions in the specialty. This procedure was considered appropriate because initial-skills courses are designed to provide training for tasks likely to be encountered during the first term of service.

A wide variety of dependent variables falling into three broad categories were examined. The dependent variables included those which were expected to be highly related to the LD index as well as those which were not. For dependent variables in each category, detailed descriptions of the source of data and measurement process are provided by Mumford et al. (1985).

The first category of dependent variables consisted of measures of student inputs. These measures included (1) students' average scores on the composite of the Armed Services Vocational Aptitude Battery which serves as the basis of the aptitude requirement for entry into the course, (2) average reading grade level as measured by the Air Force Reading Abilities Test, (3) the average academic motivation of students in the course as indexed by the average number of difficult high school courses completed, (4) average educational level of students in the course, (5) educational preparation as indexed by the average number of recommended high school course prerequisites completed by students in the course, and (6) the average age of students in the course.

The second category of dependent variables consisted of measures which represent outcomes of training. These measures included (1) average final course grades for students in the course as indexed by the average score on end-of-block written tests, (2) average number of hours of special individualized assistance (SIA) provided students by course instructors, (3) the number of academic counseling sessions provided students, (4) the number of nonacademic counseling sessions, (5) washback time as indexed by the average number of retraining hours provided students, and (6, 7) the academic and nonacademic student attrition rates for the course.

The third category of dependent variables consisted of measures of properties of the initial-skills course and are described as course content variables. This category included (1) course length (in hours), (2) course diversity as reflected by the number of different units of instruction, (3) expert's average rating of the abstract knowledge requirement of the course,

(4) the expected student attrition rate for the course, (5) the number of students per instructor, (6) the average number of months of instructional experience of the instructors assigned to the course, (7) average ratings of the quality of instruction provided by course instructors, (8) manning requirements as reflected by the availability of a selective reenlistment bonus for the specialty associated with the course, (9) length of academic day (in hours), (10) the number of instructional aids employed per instructional hour, (11) the percentage of training hours devoted to hands-on instruction, (12) the frequency of formal feedback per hour of instruction, (13) practice or the number of hours devoted to a unified body of material, (14) student flow or total number of students passing through the course per year, and (15) the reading difficulty of course materials.

Analyses undertaken to examine the construct validity of the LD index were straightforward. Scores on the various student input, training outcome, and course content variables were correlated with the LD index. With regard to analyses involving student input variables, it is important to note that the LD index for the job specialty associated with a course was assumed to be applicable to all students in the course. Once the correlational analyses were completed, the pattern of relationships indicated by the correlational data were examined to evaluate the discriminant and convergent validity of the LD index.

Results and Discussion

Table 1 presents the mean and standard deviation of each dependent variable and its correlation with the LD index. Examination of the student input variables indicates that all observed correlations were statistically insignificant. Because the LD index is a measure of a job property and the student input variables are measures of personnel attributes, this outcome was expected. The fact that this expectation was confirmed by the results of the present analyses lends support to the discriminant validity of the LD index.

All observed correlations for the training outcome variables were statistically insignificant. This outcome was also expected. As a result of efforts to develop the covariance-structure model of training, we have gained some insight into the complex interrelationships that combine to influence training outcomes. Largely because of the instructional systems design process which uses information concerning job tasks as one basis of course design, the LD index is conceived of as having a primal influence on course content. However, numerous student input and course content variables combine to influence training outcomes. Consequently, the relationships between the LD index and training outcomes are indirect being moderated by several other variables.

The strongest relationships produced by the LD index were with various course content variables. This was expected because as previously indicated, the design of initial-skills training is guided by job content. For instance, because the LD index represents learning time, it was expected that a positive relationship would exist between the LD index and course length. The fact that a moderate positive correlation was observed

for these two variables tends to argue for the convergent validity of the index. Moreover, it would be expected that the LD index would be positively related to indices of course subject matter difficulty to the extent that it is an intrinsically meaningful index. The moderate positive relationships observed between the LD index and course diversity, abstract knowledge requirement and expected attrition rate support this expectation.

In addition to these relatively direct relationships, it was expected that the LD index would yield a number of more diffuse relationships. For instance, it might be expected that fewer students per instructor and more experienced instructors would be a means of compensating for task learning difficulty. The observed relationships between the LD index and these two variables lends support to this expectation. Further, it might be expected that the overall quality of instruction would be low for tasks of high learning difficulty. Again, the observed relationship between these two variables supports this expectation. The positive relationship between the LD index and manning requirements may be attributed to the fact that specialties higher in learning difficulty generally require training which is highly valued in the civilian work place. The loss of military-trained technicians to the civilian labor market would, in turn, lead to a greater demand for personnel.

Table 1. Means and Standard Deviations of Dependent Variables and Bivariate Correlations with the Learning Difficulty Index

VARIABLES	MEAN	STANDARD DEVIATION	r
Student inputs (N = 5,970 students)			
Aptitude Composite Score	68.800	16.30	.046 NS
Reading Grade Level	11.400	1.00	.064 NS
Academic Motivation	39.000	13.40	.069 NS
Education Level	2.180	.45	-.012 NS
Educational Preparation	1.670	.94	-.118 NS
Age	20.100	2.20	.035 NS
Training Outcomes (K = 48 courses)			
Final Course Grade	85.200	7.61	.076 NS
SIA Time	6.560	15.40	.048 NS
Academic Counseling	1.450	3.47	.043 NS
Nonacademic Counseling	.170	1.56	-.009 NS
Washback Time	11.100	51.50	.066 NS
Academic Attrition	.028	.16	.014 NS
Nonacademic Attrition	.004	.06	-.002 NS
Course Content (K = 48 courses)			
Course Length	413.900	309.30	.501 **
Course Diversity	54.800	43.30	.536 **
Abstract Knowledge Requirement	2.420	.98	.467 **
Expected Attrition Rate	.096	.03	.450 **
Student-Instructor Ratio	9.100	4.80	-.556 **
Instructor Experience	32.500	14.70	.591 **
Instructor Quality	2.500	.15	-.314 *
Manning Requirements	.440	.49	.423 **
Day Length	.480	.44	-.033 NS
Instructional Aids	.270	.10	.161 NS
Hands-on Practice	.410	.13	-.186 NS
Frequency of Feedback	.340	.12	-.191 NS
Practice	8.520	3.17	-.074 NS
Yearly Student Flow	1943.600	2662.50	-.200 NS
Reading Difficulty	10.980	.63	.031 NS

** - Probability $\leq$ .01 that observed r is a random deviation from a population r of zero.
* - Probability $\leq$ .05 that observed r is a random deviation from a population r of zero.
NS - Probability $\geq$ .06 that observed r is a random deviation from a population r of zero.

Conclusion

In reviewing the results of these analyses, it seems reasonable to conclude that the occupational learning difficulty index displays construct validity. The index was found to be moderately related to both course length and a number of indices of course subject matter difficulty lending support to the convergent validity of the index. Moreover, the finding that strong relationships did not exist between the LD index and a number of student input and training outcome variables tends to support the discriminant validity of the index. Overall, the evidence produced by these analyses appears to indicate satisfactory construct validity for the occupational-level index of learning difficulty.

References

Burtch, L. D., Lipscomb, M. S., & Wissman, D. J. (1982, January). Aptitude Requirements Based on Task Difficulty: Methodology for Evaluation (AFHRL-TR-81-34). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. Psychological Bulletin, Vol. 56, pp. 81-105.

Hendrix, W. H., Ward, J. H. Jr., Pina, M. Jr., & Haney, D. L. (1979 September). Pre-enlistment Person-Job Match System (AFHRL-TR-79-29). Brooks AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory. (AD-A078 427).

Mumford, M. D., Weeks, J. L., Harding, F. D., & Fleishman, E. A. (1985). An Empirical System for Assessing the Impact of Aptitude Requirement Adjustments on Air Force Initial-Skills Training (AFHRL-TR-). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. (In press)

Weeks, J. L. (1984, November). Occupational Learning Difficulty: A standard for determining the order of aptitude requirement minimums. (AFHRL-SR-84-26). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. (A417 410).

News about CAT in the German Federal Armed Forces (GFAF)

Wolfgang Wildgrube
Psychological Service, Bonn, West Germany

Introduction

At the last annual MTA-conference 1984 in Munich, a workshop about "CAT in Germany" had presented the first activities and experiences used computerized testing in the GFAF (named: CAT I). Since 1983 the GFAF evaluated two experimental installations for computerized testing at the recruiting centers in Munich and Hannover choosing the practical approach by conventional testing (Wildgrube, 1985 b; see also the contributions of Angermüller and Kulling).

Meanwhile CAT made an important step forward in the GFAF. First a lot of paperwork to prepare computerized testing is accomplished so that in 1987 CAT systems can be operational at the four recruiting centers for volunteers and the recruiting center for officer candidates compiling the conventional procedure to computer application. Second a major change is planned at the recruiting centers for draftees. The final goal will be to accomplish medical examination and psychological testing - supplemented by psychological counseling - at the same day so that each draftee knows, after one day at the recruiting center, the date, location, and unit for the time in the service respectively for his beginning basic training. Therefore individualised testing by computer is necessary for this one day examination for draftees. The corresponding pilot project started in January 1985 in Hannover using one of the CAT installations.

Aptitude Classification Battery

In the GFAF the Aptitude Classification Battery (EVT - German abbreviation) is in use as the standard entrance examination for draftees and volunteers (similar the ASVAB; officer candidates have a special test battery). Parallel to the CAT developments goes a major revision of the EVT battery, so that the GFAF starts in January 1986 with the following revised test battery:

Figure Matrices Test (FMT)	20 Items/8 Alternatives	18 Min.
Word Relation Test (WBT)	20 Items/5 Alternatives	4 1/2 Min.
Arithmetic Reasoning Test (RT)	20 Items/Input of the Results (paper/pencil for notices)	14 Min.
Spelling (Orthographical Test; RST)	60 Items/4 Alternatives	12 1/2 Min.
Mechanical Ability Test (MT)	20 Items/5 Alternatives	13 Min.
Electrotechnical Comprehension Test (EKT)	20 Items/5 Alternatives (Pretest after RT: 8 Items	20 Min. 6 Min.)
Reaction Test (RP)	64 Items/6 Alternatives Input of the Results	3,41 Min.
Radio Test (FT)	150 Items/3 Alternatives Input of the Results	3,30 Min.
Signal Test (SigT)	18 Items/4 Alternatives Input of the Results	3 Min.
Doppler Test (DopT)	20 Items/3 Alternatives Input of the Results	4 Min.

After the six conventional subtests, up to now carried out by paper pencil, follow four special tests presented by maschines ("apparative" tests). This fixed sequence of subtests will be used during the routine application of the EVT by computer. Changes are possible at any time by interrupts of the proctor, for example omitting of the signal test or stopping after the electrotechnical comprehension test.

New Hard- and Software

Concerning the rapid developments and changes in the area of Personal Computer the definition of new hardware and furthermore, software for CAT was necessary (named: CAT II). Besides the change from PC with 8 bit processors to PC with 16 bit processors the new equipment will contain as an expansion of CAT I further the four "apparative" tests, namely reaction test, signal test, doppler test, and radio test, so that the whole Aptitude Classification Battery can be presented by computer.

There was conformity about that standard hardware and software were not sufficient for CAT at fixed locations/recruiting centers in the GFAF. At the end of the last year a booklet with the detailed requirements was worked out in cooperation between the GFAF and the German firm ZAK. The new hard- and software will be delivered in December 1985. One installation of the new equipment is assigned for the recruiting center of

draftees in Munich, the other one is a twin-configuration for the recruiting center in Hildesheim which should be delivered in Spring 1986.
In addition to the actual developments by the firm ZAK various firms present at the moment different concepts and corresponding financial estimations to realise the requirements of the GFAF concerning CAT for example in fifty recruiting centers for draftees.

The new CAT equipment has the following characteristics:
- A local area network (LAN) is in use with 15 work stations for examenees applicable different test batteries at each station.
- The testing session will be monitored at a central place by an IBM AT with harddisk. A second IBM AT - also in the LAN linked - is located in a seperate room and will be used for input of personnel/biographical data and for output of results and furthermore the second central place is prepared as back up for the master.
- Back up-arrangements are prepared at different levels, e.g. the central places for monitoring (2 IBM AT), the work stations (15 stations available), power interrupt (additional equipment and software tools), test results (internal twin storage at disk). There is no break allowed longer than 30 minutes as well as the loss or the repeat of a whole subtest.
- A work station can be described by these characteristics:
  + a white screen with black items, 768 x 1024 pixels;
  + a headset for voice output;
  + a special keyboard with a part for the compiled paper pencil tests (ten digits and the green and red function keys) and with four special parts for the "apparative" tests (e.g. reaction test).
- At the central place of the CAT station monitors the proctor the test session, using different menues. A maximum of four tasks are active so that an overview about the state of testing at any time is possible.
- The second proctor sits in the seperate room, types in the personnel data (at the moment the system provides 18 menues), or starts the output by a matrix printer (the different output is tailored for the requirements in the recruiting center, at subtest-, person- or item-level). The reference criterion for all data is the personnel-id-number (similar the social security number) so that all data (personnel data, biographical data, test data/results) are stored in a data base at the end of the whole psychological examination. The possibilities for transfer to a mainframe computer is considered, at subtest-level online and at item-level via tapedrive for follow up studies.
- The testing session starts for each examenee with an introduction or learning phase. There is the chance to get practice in the use of the keyboard and to learn the kind of testing. If an example solved or typed in incorrectly the program presents the example item once more. This phase and the example items before each subtest are supported by voice

output via headset. The whole text (without the items within the subtests) is prepared by deltamodulation, stored at the harddisk at each work station, and then monitored (screen and voice) by program during the test session.

CAT Center

At the recruiting centers the testing station is only installed for the application of the tests. After given in the password the proctor starts the session using the different menues, while a special password is necessary for data handling or for use of utilities. All aspects of change of the items or the itempool or modification of the software or the testing procedures will be carried out in the CAT center which is located at the Federal Armed Forces Office in Bonn.
Until to the end of this year the following hardware will be delivered: 1 central place (IBM AT), 1 work station, printer, plotter (for graphical items), tape drive (for data transfer to mainframe at item level). Furthermore the essential software will be prepared for Bonn:

- Source codes of the whole application software.
- Compiler for Basic, FORTRAN, Pascal, C (the greatest part of the software is written in C).
- Tools used by the firm ZAK for the software developments.
- Item editor for the input of graphical/nongraphical items, instructions, voice modules, and for storage in an itempool respectively data base.
- Item editor for assembling different subtests, modifying testing procedures, inserting instructions, and creating new test batteries.
- Utilities for data handling and for the management of data in a base.
- Statistic software for simple analyses at the personal computer.

All changes and modifications are prepared in the CAT center in Bonn and the floppies containing the newest version are distributed to the recruiting centers for the daily routine testing. In addition the CAT center is the link between GFAF and firm so that all hardware troubles first are reported to the CAT center. So the CAT center will be the central place for all aspects of CAT in the GFAF concerning hardware, software, itempools, data of the examenees, as well as for the scientific evaluation.

Experiences and Results

The two pilot installations in Munich and Hannover (CAT I) are in use daily and the CAT results are applied to selection and classification decisions. Now the data transfer is available from CAT to mainframe computer via tape drive for different statistical analyses.
Besides the computerized testing information has been collected by an ad-hoc-questionnaire. Here are some results from selected questions collected in Hannover (also classified by the four educational levels) and Munich.

	Hannover (N=223) level of education					Munich (N=175)
	1	2	3	4	T	T
1. When I entered the testing room and saw the testing equipment, I was curious as to what expected me						
1. I was very curious	2	53	78	32	165	136
2. I was not so curious	-	22	18	9	49	32
3. I was not curious at all	-	2	5	2	9	7
14. Comparing the computerized test version to the paper and pencil test, I think						
1. that I like the computerized form better	1	61	80	32	174	134
2. that I like the paper and pencil form better	1	11	15	8	35	33
15. I often play video games (e.g. "Star War")						
1. yes	1	14	17	8	40	39
2. no	1	59	80	31	171	129
16. I have experience with home computers						
1. yes	-	6	22	8	36	21
2. no	2	69	75	31	177	149
17. I participated in this test with pleasure						
1. if yes, why	2	70	92	34	198	145
2. if no, why not	-	5	3	4	12	18

The results indicate a high acceptance of computerized testing respectively this kind of non-group testing by computer. The examenees prefer the CAT application even if they have no experiences with home computers or video games. So will be soon computerized testing without any problems similar the conventional paper pencil testing.

CAT offers the chance for recording more data as well as paper pencil testing, so that more detailed analyses can be made. One aspect is the item solution time which is recorded for each item. But at time new models in testing theory and basic research are necessary to interpretate this time-based data in addition to the ability parameter.
An other important point concerns the difference in the test scores during a day period shown in the following table.

Test (time)	Scores morning	noon	p	Time used per subtest (seconds) morning	noon	p
WAT (11.00)	13.86	13.01	.002	234	246	.000
FDT (11.00)	14.46	13.76	.031	505	510	.504
RT (11.00)	9.99	9.13	.008	826	823	.690
MT (11.30)	11.59	10.70	.001	712	711	.890
RST (12.00)	28.10	26.00	.004	172	175	.016
EKT (12.00)	6.63	5.55	.000	552	561	.202

There are significant and relevant differences between the two groups (sample sizes approximately equal) in the scores of all six subtests. Remarkable are the significant values for the two subtests "word analogy" and "spelling" concerning the time used per subtest. The table above shows for the other subtests very similar time for solving the items in a subtest, while the scores are different. Further research is needed for this aspect so that different norms will be used for the morning and for the noon session if necessary.

Wildgrube, W. (1985, a): Computerized Testing in the German Federal Armed Forces (FAF): Empirical Approaches; in: Weiss, D.J. (Ed.) Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference, Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnesota, April 1985, p. 353 - 359.

Wildgrube, W. (1985, b): Computerized Adaptive Testing (CAT) in Germany - General Topics -; in: BMVg - P II 4 (Eds.), Proceedings - 26th Annual Conference of the Military Testing Association, München, Fed. Republic of Germany, 5 - 9 Nov. 1984, coordinated by the Psychological Service of the Federal Armed Forces - BMVg, P II 4, Bonn, 1985, p. 151 - 156.

EFFORTS TOWARDS THE IMPROVEMENT OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

Deirdre J. Knapp, Rebecca M. Pliske, and Richard M. Johnson
U.S. Army Research Institute for the
Behavioral and Social Sciences[1]

The Computerized Adaptive Screening Test (CAST) was designed by the Navy Personnel Research and Development Center (NPRDC) under the sponsorship of the Army Research Institute (ARI) to provide a prediction of prospects' Armed Forces Qualification Test (AFQT) scores at recruiting stations. This paper briefly discusses the development of CAST and summarizes current efforts to enhance its utility.

Background

Armed Forces applicants fall into certain mental categories on the basis of their AFQT scores. AFQT, which is intended to be a measure of trainability, is derived from a linear combination of subtest scores (i.e., WK, AR, PC, and one-half of NO) on the Armed Forces Vocational Aptitude Battery (ASVAB). In the Army, individuals who score at or above the 50th percentile (mental categories 1, 2, and 3A) are eligible for special options and benefits such as the 2-year Enlistment Option and the Army College Fund. Applicants who score between the 31st and 49th percentiles on AFQT (mental category 3B) qualify for enlistment but are not eligible for special options. Lastly, those individuals who score below the 31st percentile (mental categories 4A, 4B, 4C, and 5) are regarded as being low priority candidates for enlistment.

It is vital that recruiters have access to information which predicts prospects' AFQT performance for several reasons. For example, recruiter's missions specify not only the number of recruits to be enlisted, but also the quality of those recruits as determined by mental category classification. Further, if a prospect appears to have virtually no chance of producing an acceptable AFQT score, the recruiter may choose to discourage him or her from further interest in the Army. The recruiter can then spend more time selling the Army to more promising prospects. Finally, if a prospect appears to be of average quality, the recruiter may not want to spend much time describing special options and benefits to the individual. On the other hand, if the recruiter does not sell the options and benefits to those individuals who are likely to be eligible for them, he or she is failing to use a powerful sales tool. Clearly, recruiters can enhance their performance if they effectively use a valid predictor of prospects' AFQT performance.

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

Description

As its name indicates, CAST is a computerized adaptive test. Adaptive tests are constructed so that they are tailored to fit each examinee. This is done by administering items which have optimal discriminability given a particular examinee's ability. This can be compared to traditional testing where all examinees respond to the same items, regardless of differences in examinee ability. Hence, adaptive tests are more efficient to use than traditional tests because a comparable amount of information can be gained from fewer test items.

The item pool for CAST was developed by researchers at the University of Minnesota (cf. Moreno, Wetzel, McBride, & Weiss, 1983) for use in the development of a computerized adaptive version of ASVAB. Researchers at NPRDC developed the software capable of administering items predictive of AFQT on the Army's microprocessor system known as JOIN.

The current operational version of CAST consists of 78 word knowledge (WK) items and 225 arithmetic reasoning (AR) items. All items are multiple choice with a maximum of five response alternatives. The items were developed using the three-parameter logistic ogive item response model (Birmbaum, 1968); thus each item has three parameters (discrimination, difficulty, and guessing) associated with it. Test items for CAST were chosen so that the discrimination parameter values would be greater than or equal to .78; the difficulty parameter values would range between +2 and -2; and the guessing parameter values would be less than or equal to .26. The ability estimates yielded by CAST are based on the Bayesian sequential scoring procedure discussed by Jensema (1977). The stopping rule is ten WK items and five AR items.

Validation Information

There are three validation efforts associated with CAST. The initial validation study was conducted at the Los Angeles Military Entrance Processing Station (MEPS) with a sample of 312 U.S. Army applicants (Sands & Gade, 1983). The correlation between CAST scores and AFQT scores was .85. The second data collection effort took place in Army recruiting stations in the midwestern region of the U.S. during the first two months of 1984 (Pliske, Gade, & Johnson, 1984). CAST scores were linked to subsequent AFQT performance via the social security numbers (SSN's) of the prospects. More specifically, recruiters recorded prospect SSN's thus allowing the researchers to locate the appropriate MEPS records. Matching records for 1,962 prospects were located and the resulting validity estimate was .80.

The most recent estimate of CAST's validity is based on data which is being collected from a sample of 60 Army recruiting stations during January through December of 1985. This sample was selected to be representative of the population of approximately 2,000 Army recruiting stations in terms of geographic location, population density, and ethnic composition. The correlation between CAST and subsequent AFQT performance, based on preliminary analyses of the first six month's of data, is comparable to those obtained in the earlier studies (r=.82; n=2,240).

Clearly, CAST is a valid predictor of AFQT performance. Of course, one problem with computerized adaptive testing is the need to have a computer standing by to administer it. Since recruiters do not always have this luxury, there remains the need to use the Enlistment Screening Test (EST), a paper-and-pencil predictor of AFQT. The initial validation information regarding EST was provided by Mathews and Ree (1982). Although their data yielded a healthy validity estimate of .83, a cross-validation of the test has not been reported. Consequently, the recruiting stations which have been providing CAST validation data over the past year have also been asked to record the SSN's and EST scores of all prospects who take EST rather than CAST. Based on six month's of data, the validity estimate is .79 (n=685). As expected, the validity of EST has been reaffirmed.

Proposed Improvements

Currently, efforts are underway to improve three specific aspects of CAST. The first aspect concerns the kind of information that the test provides to the recruiter. The second and third aspects involve the test's item pools and stopping rule. At the present time, CAST provides bar charts that give information about performance on the WK and AR subtests and the examinee's predicted AFQT percentile score. There are two fundamental problems with providing only point predictions of AFQT scores to recruiters. The first problem is a function of the statistical naivete' of most recruiters. Because the great majority of recruiters do not understand the concept of correlation, they do not adequately understand the nature of the point prediction that they are given. Hence, recruiters complain that predicted and actual AFQT scores often fail to be exactly the same. A second problem with the use of point predictions concerns the way in which recruiters utilize information from CAST. As indicated earlier, recruiters are primarily concerned with the prospects' subsequent classification into one of three groups of mental aptitude categories. Given these considerations, it seems that recruiters would be better served if CAST provided output that reported the odds associated with a given prospect falling into each of the three critical mental categories.

Two approaches to category prediction are being studied using data from the on-going validation effort described above. The first approach models the strategy that recruiters probably follow. That is, one uses the point prediction, which is based on a regression model, to determine the mental category to which an individual will likely belong. The second approach is based on classification analysis. In contrast to regression analysis, classification analysis provides subtest weights that optimize category, rather than point, predictions. Table 1 shows the percentage of cases which were classified into each of three categories on the basis of these two approaches. A comparison of the two approaches reveals that they differ with respect to where their prediction errors occur. Both approaches are good at identifying individuals who fall into categories 1-3A (Approximately 75% of the people who are predicted to be in 1-3A actually are in 1-3A). Classification analysis, however, is much better than regression analysis at identifying individuals who are in categories 4A and below (80% versus 55% accurate prediction). This advantage is at the expense of a somewhat poorer ability to identify prospects who are in category 3B.

TABLE 1

Percentage of Cases Classified
Into Critical ASVAB Categories

		PREDICTED AFQT CATEGORY* 1-3A	3B	4A AND BELOW
ASVAB	1-3A	76/77	23/18	1/5
AFQT	3B	25/27	60/38	15/35
CATEGORY	4A and below	4/4	41/16	55/80

* Regression analysis to left of diagonal; classification analysis to right of diagonal.

Regardless of which method of category prediction is used, the information provided can be presented in a way that would be logical to recruiters. For example, CAST software could be changed to report the probabilities associated with an examinee falling into each of the three critical categories. The recruiter could then compare the odds, and make an informed judgment concerning his or her subsequent course of action. The important point here is that presenting CAST results in such a fashion would make it clear to recruiters that, although those results can be very useful, they are not infallible.

Turning to the subject of CAST's stopping rule, the questions to be asked are twofold. First, assuming that the stopping rule is to be based on number of items, what is the optimal number of subtest items to administer? Second, would it be better to base the stopping rule on the precision of the ability estimate rather than on the number of items being administered? Data from the latest validation study have been used to examine the first question. The version of CAST which is being used in the 60 experimental recruiting stations administers five more items per subtest than the operational version of the test. It also records response time so that the time it takes to administer various subtest length combinations can be compared.

Validity estimates and average completion times were computed for six subtest length combinations (5, 10, and 15 WK items; 5 and 10 AR items). Validity coefficients ranged from .79 to .85. Completion times ranged from just over 10 minutes to just over 18 minutes. Given that the validity estimate associated with the current stopping rule is .82 and the average completion time is a little over 12 minutes, it appears that an increase in subtest length would not be justified. A substantial increase in completion time is required for even a small increase in validity.

With an adaptive test, an ability estimate and the variance associated with that estimate is computed each time an examinee answers a new test item. Rather than stop the subtests after a given number of items are administered, the test software can be altered to end the subtests once the variance estimate has dropped to a given value. The advantages and disadvantages of al-

tering CAST to rely on the latter type of rule need to be evaluated carefully. Also, the determination of an optimal variance criterion would require further study.

The two item pools currently contained in CAST will be expanded within the next 2-3 years. Given the extensive use of the test, it is important that the item pools are large enough to prevent the frequent recurrance of particular test items. The possibility of developing items which will provide optimal discrimination at the critical AFQT cutpoints is also being considered.

Closing Remarks

CAST is a very good test which we are seeking to make even better. At the present time, our efforts are primarily aimed at changing the software to yield information that will be of the greatest use to recruiters. Our plans over the next couple of years are aimed at insuring the continued integrity of the test itself.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds) Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley.

Jensema, C.G. (1977). Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1, 111-120.

Mathews, J.J. & Ree, M.J. (1982). Enlistment Screening Test Forms 81a and 81b: Development and Calibration (AFHRL Report No. 81-54). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1983). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests (NPRDC) Report No. 83-27). San Diego, CA: Navy Personnel Research and Development Center. (NTIS No. ADA 131683).

Pliske, R.M., Gade, P.A., & Johnson, R.M. (1984). Cross-Validation of the Computerized Adaptive Screening Test (CAST). Alexandria, VA: U.S. Army Research Institute.

Sands, W.A., & Gade, P.A. (1983). An application of computerized adaptive testing in U.S. Army Recruiting. Journal of Computer-Based Instruction, 10, 87-89.

DEVELOPMENT OF A PORTABLE COMPUTERIZED PERFORMANCE TEST SYSTEM

R. S. Kennedy, W. P. Dunlap,
R. L. Wilkes, & N. E. Lane
ESSEX Corporation
Orlando Florida

Abstract

The ethics and pragmatics associated with developing an automated performance test system to study the effects of various treatments make repeated measures in small groups of subjects the customary research paradigm. In such cases, test stability, reliability, and factor structure take on extreme significance. In a Navy program 80 percent of 150 tests failed to meet minimally acceptable psychometric requirements. Recent findings with our battery show: acceptable psychometric properties in terms of both differential stability and reliability for both the long and the short battery; two factors available for the 7.5-minute test battery; four for the 15-minute battery and correlation with the WAIS. The factorial richness of the battery is adequate and goes beyond the factors that can be conveniently measured by more traditional paper-and-pencil tests into motor speed dimensions that may have important practical implications for assessment of concurrent functional capacity. Both factorial richness and correlation with the more global cognitive capacity construct, IQ, might be improved by the inclusion of subscales indexing verbal and arithmetic abilities, however, adding factors is not without penalty. The trade-offs of these issues (testing time, factor structure, stability, stabilization time, and reliability) are discussed. About a dozen validation studies are presently ongoing. What remains is to demonstrate functional validity in the detection of human functional capacity deficits in a real-world setting. This should be our primary mission in subsequent work.

INTRODUCTION

Exotic work environments often include factors (i.e., weightlessness, motion, fatigue, etc.) that disrupt performance. Furthermore, these settings are typically populated by limited numbers of highly critical workers. Kennedy and Bittner (1977) have observed that two connected concerns associated with the measurement of performance under such conditions are the lack of sensitive tests and a general unwillingness to expend the time and effort necessary to standardize such tests. A program designed to evaluate performance measures (PETER) was undertaken by the Naval Biodynamics Laboratory, New Orleans, LA (Kennedy & Bittner, 1977; Bittner & Carter, 1981; Kennedy, Bittner, Harbeson, & Jones, 1982), and more than 150 performance tasks have been examined for suitability in repeated measures research. Detailed descriptions of the evaluation process and task metric selection criteria may be found in Bittner, Carter,

Sponsored by NASA No. 9-16982 and NSF No. 00559

Kennedy, Harbeson,, and Krause (1984). A listing of 30 tests which survived this test and evaluation process appears in Bittner et al. (1984), but for the most part the tests that were studied were paper-and-pencil tests. The easy availability and economy of portable high speed computers suggests that innovative methods for automated data collection and analysis must be explored. Features that recommend microbased testing systems include capabilities for fully automated test battery administration and data storage, as well as portability and reduced size and weight. Automated and portable microprocessors capable of administering and storing performance measures and responses provide the obvious vehicle. The purpose of this report is to provide a descriptive overview and a brief prospectus of the Automated Performance Test System (APTS), and report our recent progress in the engineering analysis of the battery in terms of stability, reliability, factor structure, feasibility, and predictive validity. Ongoing validation studies which examine sensitivity to treatments are being recounted elsewhere at this meeting (Johnson, Kennedy, Merkle, Smith, & Bittner, 1985).

RECENT TEST DEVELOPMENTS

Preliminary Study

Method and Analysis. The "best" six tests from the PETER program (Bittner et al., 1984) were programmed on a portable microprocessor and administered along with tests in their original paper-and-pencil formats. Twenty-three Casper College male and female students were tested over four replications on a 6.0 minute computerized battery. The group means, standard deviations, and 4X4 intersession correlation matrices were calculated for each task in each testing mode. Task group means and standard deviations were examined across sessions for evidence of task stabilization. Intersession correlations were assessed for evidence of task differential stability. Rapid stabilization was expected since at least theoretically comparable practice was received within both modes of testing.

Results and Discussion. The data showed that all tasks in both modes give good evidence of stability by the fourth session, with high reliability efficiencies (r >.85) for 3 min. of testing. Improvement, averaged across all tasks from sessions 1 to 4, was approximately 20%. The amount of improvement for paper-and-pencil testing (22.4%) was, in general, comparable to the amount of improvement demonstrated in the microbased testing mode (19.3%). Typically, paper-and-pencil testing produced higher scores across test sessions relative to microbased testing; however, from the data the acquisition curves for both modes are strikingly similar. Furthermore, the task standard deviations provide good evidence that none of the tasks in either mode has reached a ceiling. Clearly, all indicators point to good and comparable metric characteristics for the paper-and-pencil and microbased versions of each task. The factor structure obtained from the analyses of computerized test versions in each of the four sessions indicates the presence of two well-identified factors in the computer battery. Factor 1 is clearly a "motor" factor, probably related to response speed; as such, it affects performance on Pattern Comparison and Grammatical Reasoning. Factor 2 is just as definitively a "cognitive" factor with its importance for various tests changing with practice. The

clarity of analyses under this constraint is encouraging. With respect to the paper-and-pencil tests, there is reason to believe that these are essentially the same factors as for computerized versions, but the computerized versions appear to stabilize earlier and to be more clearly defined. It should be noted that in both the microbased and paper-and-pencil analysis a possible third factor gave indications of emerging. The nature of the factor is unknown; however, the "automaticity" of responses characteristic of well-practiced skills (Ackerman & Schneider, 1984) is a likely potential explanation. Also, a significant general factor "g" may become evident within both modes of presentation. These findings are described in detail elsewhere (Kennedy, Wilkes, Lane & Homick, 1985).

Stabilization Study

Method and Procedure. Thirty-one male and female college students were recruited for participation. Prior to testing, subjects received a brief introduction to the purpose of the study and were advised regarding the general procedures associated with data collection. Ten (10) paper-and-pencil batteries and ten (10) microbased batteries were administered per subject, and the subjects were also tested individually with the Wechsler Adult & Intelligence Scale (WAIS). Six of the tests previously recommended above as a "mini-battery" for environmental research were included for examination. Four additional tests were also studied. The group means, standard deviations, and intersession correlation matrices were calculated for each individual paper-and-pencil and microbased test. Group means and standard deviations were examined for evidence of test stabilization, and intersession correlations were assessed for evidence of differential stability. Task definition (magnitude of r after stabilization) was determined directly and then adjusted according to the Spearman equation for test length and called average stabilized "reliability-efficiency" (to a 3-minute base). Predictive validity was assessed by comparison with an individually administered test of intelligence. Factor analyses used the principal factors method with squared multiple correlations as community estimates, followed by normalized varimax rotation.

Results and Discussion.

o Stability of Means - The overall impression of the 10 tests was that continued improvement occured over all tests but is slowed down in those tests considered stabilized. The Sternberg test appears to stabilize by Trial 4. The Preferred Hand Tapping and Non-preferred Hand Tapping tasks stabilize rather late in practice, by about Trials 7 or 8, whereas, the Two-Hand Tapping task appears stable by Trial 4. The Pattern Comparison test stabilizes very rapidly, by at most Trial 3; whereas the Manikin test stabilizes later at Trials 6 to 7, as does Code Substitution. Gramatical Reasoning stabilizes rapidly, by Trial 3, and Reaction Time by Trial 5. The Landolt C test of dynamic visual acuity did not appear to stabilize over the 10 test days. There appears to be a strong learning component in this test as it is presently structured on the computer; thus it's stability is insufficient to be retained in the test battery in its current form. Clearly, the acuity test is a candidate for improvement in future versions of the performance battery.

o Standard Deviations - The standard deviations are constant and give no evidence of ceiling effects.

o Differential Stability - The Sternberg, the three Tapping tasks, Pattern Recognition, and Manikin reach apparent differential stability by Trial 3. Code Substitution, Gramatical Reasoning, and Reaction Time reach differential stability somewhat later, but apparently by Trial 6. The average reliabilities of these tests are all quite high. The intercorrelation matrix for the Moving Landolt C, on the other hand, does not give any indication of reaching differential stability.

o Factor Analysis - We recognize the limitations of performing a factor analysis with such a small sample, but are somewhat encouraged by the good stability and high reliability of the tests and plan for these results to be advisory. Factor I loads on Pattern Comparison and Code Substitution of both computer based and paper-and-pencil versions, and also loads on the Manikin and Fitts Histogram test which did not have dual versions. Factor II appears to be a Motor Speed factor because it loads on Reaction Time and the three Tapping tasks as well as the Sternberg, but loads on none of the paper-and-pencil tests; therefore, Factor II may represent a construct that can be measured via computerized testing but not by standard paper-and-pencil tests. Factor III is probably best thought of as a Motor Control factor because it loads on Tapping tasks and Spoke and Aiming. Factor IV is a pure Gramatical Reasoning factor, loading on both forms of this test.

o Correlations with WAIS - The microcomputer based tests clearly correlate most strongly with Performance IQ and less strongly with Verbal IQ. The strongest simple correlation between the computerized tests and Full Scale IQ was for Gramatical Reasoning, although Non-preferred Hand Tapping was fairly high. The R squared values indicate that a substantial proportion of Performance IQ variance can be predicted from the computerized battery, but may also suggest that a more verbal subtest would improve the relation to Full Scale IQ.

o Paper and Pencil Tests - These tests essentially replicated the microprocessor based tests and a more complete description of these findings appears in Kennedy, Dunlap, Jones, and Wilkes (1985).

o General - All but one of the microcomputer tests could be recommended for a mid-ranged (<10 min.) battery. Based on the factor analysis, we would suggest Pattern Comparison, Sternberg, Tapping, and Grammatical Reasoning. We would also propose that each test be administered twice as long in order to improve the reliability and thus afford an opportunity for improved sensitivity.

Conclusions

The philosophy of our approach to performance test development involves three different phases. The first is to deal with only tests or tasks that can be shown to be psychometrically sound. This requires that we demonstrate stability of means and standard deviation within few administrations, and most important, that differential stability, the symmetry or constancy of trial-to-trial intercorrelations, be shown to

occur quickly and at high values. The second phase is to show that the battery has factorial multidimensionality and that the subscales cross-correlate with earlier performance tests and other recognized instruments of ability. Finally, it is necessary to demonstrate and document sensitivity to factors known to compromise performance potential in laboratory and ultimately real world situations.

The cross-correlation between the Portable Human Assessment Battery and the WAIS IQ measures was particularly interesting in that a substantial portion of the performance subscale variance could be accounted for by this self-contained self-administered short battery of computerized tests. If arithmetic and verbal subtests were implemented we would be able to substantially improve the correlation with full-scale IQ. Of very great interest were the strong relations shown between Nonpreferred Hand Tapping and the verbal WAIS subscale. This was the second highest intercorrelation with the full-scale IQ. Jensen and Munro (1979) hypothesized that complex reaction times should be better predictors of general intelligence (g) than simple reaction times; and certainly this procedure is borne out by the present data, in that the Sternberg, a complex motor response task, has a higher correlation that simple reaction time with the WAIS IQ. Furthermore, Jenson and Munro (1979) found that motor speed showed as strong a relation as complex reaction time to g, an unexpected finding, but one that might relate to the strong relation between tapping and IQ in the present results. A comparison with other mental tests is in order (e.g., Armed Services Vocational Aptitude Battery (ASVAB). The extreme length of the ASVAB may be thought to improve the stability of its factors, but actually the second session of the ASVAB may be no more stable than Session 2 of the computer tests we reported, the latter taking 15 min (cf., McCormick, Dunlap, Kennedy & Jones, 1985). The Wechsler Adult Intelligence Scale (WAIS) takes 1-2 hrs. to administer and purports to sample two factors; these may or may not overlap with the four of the computer battery we report. In future studies in this program we will attempt to anchor our tests against these other better-known tests. For example, if only the computer tests are considered it may be possible to sample two motor and three cognitive factors, each within 6 min. and with reliabilities greater than r = .70% for a 3-min. base.

Based on the data reported above, we believe that the following four points are arguable and we would like to offer the following hypotheticial situations for speculation. (1) A Job Sample representing real-world work is likely to take > 100 hours to reach stability, and if a single (composite) score (e.g., correct detection) based on 60 min. of testing would be used to characterize performance, it is unlikely that such a score would have retest reliability greater than r = .60, and it might be lower. If many scores are broken out (e.g., hits, RMS error, miss distance), the individual reliabilities are likely to be lower than r = .30. Alternatively, the microprocessor that I have used would probably take < 1 hour (probably 30 min.) to reach stability on the test(s), and this total score is likely to have retest reliability (r > .90) for 12 min. of testing, as are each of the subtest scores. (2) High reliability does not assure sensitivity, but lack of reliability assures insensitivity. Most tests advocated for use in unusual environments or with toxic substances have neither been checked for reliability nor stability. (3) Our microprocessor-based battery total score correlates well with global

measures of intelligence. The single best predictor of job performance in all military jobs is a global measure of intelligence. Our battery probably shares considerable variance with military jobs and job performance. (4) After stability of performance on a Job Sample and stability of our battery's performance, if both are corrected for attenuation, a large proportion of the one hour test (perhaps 80%) would be shared with the 12 min. test, although the former may take 250 times as long to stabilize.

REFERENCES

Ackerman, P. L., & Schneider, W. Individual differences in automatic and controlled information processing. Urbana, IL: University of Illinois, Department of Psychology, 1984. (Report No. HARL-ONR-8401)

Bittner, A. C., Jr., & Carter, R. C. (1982). Repeated measures of human performance: A bag of research tools. (Research Report No. NBDL-81R011). New Orleans, LA: Naval Biodynamics Laboratory (NTIS No. AD A113954)

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1984, in press). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 112 measures. Perceptual and Motor Skills. (NTIS No. AD A152317)

Bittner, A. C., Jr., Smith, M. G., Kennedy, R. S., Staley, C. F., & Harbeson, M. M. (1984). Automated Portable Test (APT) System: Overview and prospects. Proceedings of "On the Use of Microprocessors in Psychology" held in conjunction with Psychonomic Society, San Antonio, TX.

Jensen, A. R., & Munro, E. (1979). Reaction time, movement time, and intelligence. Intelligence, 3, 121-126.

Johnson, J. H., Kennedy, R. S., Merkle, P. J., Smith, M. G., & Bittner, A. C., Jr. (1985). Microcomputer-based field testing for human performance assessment. Paper presented at the 27th Annual Conference of the Military Testing Association, San Diego, CA.

Kennedy, R. S., & Bittner, A. C., Jr. (1977). The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity Enhancement: Personnel Performance Assessment in Navy Systems. San Diego, CA: Navy Personnel Research & Development Center. (NTIS No. AD A056047)

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1982). Television-computer games: A "new look" in performance testing. Aviation, Space, and Environmental Medicine, 53, 49-53.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., & Wilkes, R. L. (1985). Portable human assessment battery: stability reliability, factor structure, and correlation with intelligence tasks. Orlando, FL: Essex Corporation. National Science Foundation, Final Technical Report 85-3.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). Preliminary evaluation of a microbased repeated measures testing system. Paper presented at the 56th Annual Meeting of the Aerospace Medical Association, San Antonio, TX.

McCormick, B. K., Dunlap, W. P., Kennedy, R. S., & Jones, M. B. (1982). The effects of practice on the Armed Services Vocational Aptitude Battery. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. Behavioral Report No. ARI-1. (Final Report on Contract No. MDA903-82-M3943).

LEADERSHIP PREPAREDNESS IN NEWLY COMMISSIONED NAVAL OFFICERS

Leanne Atwater
Alice Crawford

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Background

The growing concern for our national security and the concomitant expansion in the size of the Navy point to an increasing need for effective military leadership as expressed by the Secretary of the Navy. While leadership in general is one of the most thoroughly analyzed concepts to be found in the research literature, military leadership has had considerably less systematic attention. Military leadership, specifically leadership training for newly commissioned Navy officers, is the focus of this paper.

Navy personnel have generally viewed themselves first and foremost as leaders. They see the need for strong leaders--those who have the initiative, courage, and knowledge to think and act in situations where military objectives may not be easily recognized--as more critical than ever before.

There is concern, however, that many of the events during the past several decades have eroded traditional leadership values. The result has been an overemphasis on management-based, theoretical frameworks designed by social scientists who do not really appreciate leadership in a military context. Articles on this topic, written by the uniformed military are common. Sarkesian (1985), for example, traces the evolution of the corporate management model back to the McNamara years, which focused on cost-effectiveness and econometrics and shifted the focus of leadership in battle to the pursuit of management goals. He feels that the negative impact of this shift was felt in Vietnam. Byron (1985), believes that in peacetime the leadership demanded in combat situations is forgotten and instead, good managers are rewarded. To summarize, many feel that the Navy is losing its military leadership capability, which will in turn affect battle readiness.

An obvious paradox exists in the Navy community in light of all the concerns voiced when it comes to what should be done to develop these ideal leaders. Prior to the inception of this research effort, the authors conducted numerous interviews with senior Navy officers and a common theme emerged. Leadership training is conducted when everything else is finished--it is the last priority--and many, given their way, would throw it out altogether.

In part, this attitude toward leadership training stems from priorities. To teach leadership, something else must go. But this attitude seems to stem also from a lack of agreement as to what military leadership really is, and how (or even if), it should be trained. This definitional problem can be remedied if leadership is viewed as consisting of three dimensions--management, interpersonal skills, and warriorism. While warriorism may or may not be trainable, the majority opinion of leaders represented in a recently published book about military leadership (Taylor & Rosenbach, 1984) is that leadership skills (i.e., managerial and interpersonal skills) can be learned. This study addresses issues relevant to "teaching" these leadership skills to newly commissioned naval officers.

Current Leadership Training in the Navy

The Navy currently provides some training for new officers to help them assume their leadership role. Each of the commissioning sources has a leadership curriculum for naval officers as part of their education and preparation. The specialty schools (e.g., Surface Warfare Officers School (SWOS)) continue the leadership training process.

Until recently, the leadership training program at SWOS was a 2-week course designed as part of the Leadership Management and Education Training (LMET) program implemented Navy-wide in 1978 in an attempt to standardize the Navy's leadership training. Since 1978 many of the LMET courses have been dropped or shortened. The reasons for this are varied, but reflect to some extent the attitude mentioned earlier that leadership training is low priority training. It also reflects a desire to shorten the overall training pipeline and to get officers into the fleet sooner.

Purpose

If officer training is to be accomplished in the most efficient manner, it is imperative that the training provided be relevant, effective and not redundant. This study set out to address two primary questions. Does training make a difference? That is, are students learning anything they feel will be relevant to their leadership positions? If so, what are they learning and where, and can leadership training be provided more efficiently?

METHOD

Sample

The sample for this study consisted of 581 newly commissioned junior officers at SWOS. Four hundred and eighty-three students were questioned on their first day of SWOS about leadership training and preparation at their commissioning source. Ninety-eight additional students were questioned before and after participating in LMET. Table 1 presents a description of the sample.

BEST COPY AVAILABLE

Table 1

Description of SWOS Sample

	First Day of SWOS (N = 483)	Pre/Post LMET Sample (N = 98)
Commissioning Source		
USNA	196	3
OCS	37	68
NROTC	250	21
Academic Major		
Science/Engineering	300	41
Humanities/Other	183	59
Average Age	23	24
% choosing surface as 1st choice community	64	55
% intending to make Navy a career		
Yes	25	18
No 10	20	
Unsure	65	62

Questionnaires

Four versions of a questionnaire were designed to assess leadership training, leadership preparedness, and self-rated leadership abilities, and to collect a number of biographic and demographic characteristics of the students (e.g., academic major, age, sex, commissioning source). Among the four versions the following topics were also measured: managerial style, achieving style as measured by the Manifest Need Questionnaire, and action vs. state orientation. These topics will not be addressed in this paper.

Questionnaires were administered in classroom settings by SWOS instructors. Administrators were given a set of instructions to read to the students before filling out questionnaires. Students were assured that all information was confidential.

RESULTS

Does Leadership Training Really Make a Difference?

An important question in this study had to do with the merit of leadership training--does it really make a difference? Results indicated that leadership training is important to preparedness. A number of different analyses formed the basis of this conclusion. First, the relationship between new students' reports of their training and preparedness were correlated. As expected, the correlations between these two indicators were quite high (.5 to .7). At least in the students' minds they see their training as related to their levels of preparedness.

Second, analyses indicated that students' perceptions of their leadership training and preparation were more related to their feelings of preparation to go to war if necessary than were their self-rated leadership abilities. Preparation to go to war is not only determined by ability; training has an impact.

Third, a number of open-ended comments provided by new students indicated that they felt they needed more leadership training, especially in the interpersonal aspects, in order to best assume their role as division officer.

Fourth, 98 students were questioned before and after LMET about their leadership training and preparedness. On the basis of paired t-tests, 14 of the 21 items measuring different aspects of leadership preparedness showed significant improvements as a function of LMET. Many items had differences greater than a standard deviation. Specifics of these differences will be discussed in a later section.

What Leadership Training is Provided?

Training at the Commissioning Sources

The level of leadership training new officers received at the three major commissioning sources was of particular concern in this study. Table 2 presents a list of the leadership training and preparation topics assessed and indicates the areas where officers felt most and least trained and prepared. In general, officers entering SWOS felt they were trained "to some extent" (the midpoint on the scale which ranged from 1, "not at all," to 5, "to a very large extent") in most aspects of leadership. The overall levels of perceived preparation were somewhat higher ($\bar{x}$ = 3.6).

BEST COPY AVAILABLE

Table 2

Leadership Training and Preparation Topics Measured and Extreme Averages By Commissioning Source

	New SWOS Students[a]						Greatest Improvement after LMET (N = 98)
	USNA		OCS		NROTC		
	Trng (N = 196)	Prep (N = 196)	Trng (N = 37)	Prep (N = 37)	Trng (N = 250)	Prep (N = 250)	
a. Making the transition from the Naval Academy, OCS or NROTC to the operational Navy							
b. Taking responsibility for a division of enlisted personnel							
c. Understanding Navy procedures and protocol			3.5				
d. Relieving the division officer in your first division officer assignment	2.8	2.9			2.7	3.0	X
e. Knowing how to motivate enlisted personnel							X
f. Performing the paperwork requirements as a division officer (PMS, PQS, etc.)	2.5	2.7	2.5	3.0	2.6	2.8	
g. Managing your time and setting priorities when you have a heavy workload	4.4	4.3	3.6	3.9	3.6	3.9	
h. Talking to a large group of people who work for you							
i. Briefing your superior, or the CO about an issue in your division							X
j. Counseling subordinates about personal matters			2.4				
k. Counseling poor performers			2.4	3.3			X
l. Handling alcohol and drug abuse problems among your subordinates						3.3	
m. Resolving conflicts among your crew members					2.7		
n. Listening effectively				3.9		3.9	
o. Managing stress (i.e., lack of sleep, disappointing your boss, overwork, conflicts)	4.1	4.2	3.4				
p. Communicating with people effectively				3.9		3.9	
q. Demonstrating concern for your subordinates							
r. Setting goals	4.1	4.2				3.7	
s. Planning work	4.1	4.1				3.5	
t. Interacting with Chiefs in your division	2.7	3.1					X
u. Rewarding and disciplining your subordinates							X

[a]Only areas with highest and lowest average levels of training or preparation are presented in this table.

The levels of training differed across commissioning sources in 18 of 21 areas. For the most part, Naval Academy (USNA) graduates felt better trained than those from Naval Reserve Officer Training Corps (NROTC) or Officer Candidate School (OCS), and OCS graduates generally felt least trained. A noteworthy exception to this was training in how to interact with chiefs in the division. In this area, NROTC graduates felt best trained and Naval Academy graduates felt they had received little training.

The aspects of leadership in which new SWOS students, across commissioning sources, felt they had received the most training and were most prepared were "managing time" and "setting priorities with a heavy workload." Naval Academy graduates, in addition, felt well trained and prepared in terms of managing stress, setting goals and planning work. Graduates of OCS and NROTC, however, felt more prepared in areas of listening effectively and communicating effectively with people than they did in setting goals and planning work.

Areas where new SWOS students felt least prepared were "relieving the division officer" and "performing paperwork requirements." Those from OCS also reported little training in terms of "counseling subordinates," although they felt prepared to some extent to do this.

Although not reflected in Table 2, when examining the levels of training and preparation on the more interpersonal aspects of leadership (e.g., knowing how to motivate subordinates, briefing superiors, counseling subordinates), the levels of training were rather low, and the levels of preparation were moderate. This was especially true for graduates of NROTC and OCS.

It appears that in general, students leaving the commissioning sources feel better prepared in terms of managerial type skills (goal-setting and planning) than in the interpersonal aspects of leadership (counseling and disciplining).

LMET Training at SWOS

The results of LMET training improve this state of affairs to some extent (see Table 2). Leadership areas which showed the greatest improvements as a result of LMET were "motivating enlisted personnel," "talking comfortably before a large group," "counseling poor performers," "interacting with chiefs in your division," "rewarding and disciplining subordinates," and "relieving the division officer."

Also of interest were two items in which perceptions of the levels of training decreased from time 1 to time 2. These areas were "understanding Navy procedures and protocol" and "performing the paperwork requirements as a division officer." It seems likely that LMET served as a realistic preview and made officers aware of the large amount they didn't know in these areas.

Self-Ratings of Leadership Ability

While the majority of students' perceptions of their leadership training differed significantly across commissioning sources, self-ratings of leadership ability in eleven areas did not differ.

The area in which all students felt most capable was in "doing whatever it takes to get the job done." Aspects of leadership in which students felt least capable were "speaking comfortably in front of a group," and "motivating subordinates to do jobs they don't want to do." In general, officers feel more able to perform managerial duties than to handle the more interpersonal aspects of their leadership role.

The leadership ability ratings of Academy graduates were somewhat surprising. While Academy graduates felt better trained and better prepared than graduates of the other two commissioning sources, they did not rate their leadership abilities any higher. This seems to dispel, to some extent, the stereotype that Academy graduates have trouble adjusting in their first tour as division officers due to an over-confidence. It also suggests that abilities are not the only factor relevant to how well prepared officers feel they are. Training does play a role.

Additional Findings of Interest

Who or What Influenced Leadership Skills?

New SWOS students were asked the extent to which classroom instruction, examples of military leaders, experience leading others, as well as things they learned before they entered USNA, OCS or NROTC influenced the leadership skills they had acquired. Academy and NROTC graduates felt examples of military leaders and experience leading people were most influential. OCS graduates felt they had been most influenced by things they learned before attending OCS. OCS graduates also tended to feel their levels of training had been low in comparison to their levels of preparation. This supports the conclusion that OCS graduates feel their leadership skills are less a function of their officer training than do those who go through NROTC or the Naval Academy. (This is not surprising when the length of these officer training programs is considered (i.e., four months for OCS as opposed to four years for USNA and NROTC).) All three groups felt classroom instruction was least influential, though they still reported it influenced them "to some extent."

Assuming Different Leadership Roles

Functioning effectively as a naval officer encompasses a number of different leadership roles. The officer must function as a leader of others, a technician, a professional, and, at times, a warrior. SWOS students were asked how well prepared they felt to assume each of these roles (see Table 3). Students from all commissioning sources, at SWOS for the first day, felt most prepared to assume the role of professional, and least well prepared to assume the role of technician. One might expect that 14 weeks of intense SWO training would improve officers' confidence in their preparation to

assume the role of technician, but the sample of students questioned on their last day of SWOS also felt least prepared to assume the role of technician when compared to the other three roles. This may be because officers are trained to manage technicians rather than to develop technical expertise. Alternatively, their training may serve to make them aware of how much there is to know.

The overall levels of preparedness to go to war if necessary were fairly high ($\bar{x}$ = 3.6). All 21 aspects of leadership training were significantly related to preparation to go to war. The training variables most highly correlated with preparation to go to war were in areas of motivating enlisted personnel, handling alcohol and drug abuse problems, resolving conflicts and planning work. These correlations were all approximately .26.

Table 3

Average Levels of Preparation to Assume Diverse Leadership Roles
Before and After SWO Training by Commissioning Source

	New SWOS Students				SWOS Graduates			
Leadership Role	USNA (N = 196)	OCS (N = 37)	NROTC (N = 249)	Overall (N = 482)	USNA (N = 5)	OCS (N = 86)	NROTC (N = 34)	Overall (N = 125)
Technician	3.4	3.2	3.1	3.2	4.4	2.9	3.0	3.0
Professional	4.3	3.9	3.9	4.0	4.2	3.6	3.9	3.7
Leader	3.9	3.4	3.7	3.8	3.8	3.5	3.7	3.5
Prepared to go to war	3.8	3.8	3.5	3.6	4.0	3.6	3.6	3.6

Academic Major and Leadership Preparedness

The Secretary of the Navy, John Lehman, and others have suggested that a strict science and engineering curriculum may be too academically narrow to provide new officers with the well-rounded education they need to be good leaders. To address this question, correlations were computed between academic major (science and engineering vs. other) and students' perceptions of their leadership training, preparation and ability. (These correlations were partial correlations controlling for commissioning source.) No significant relationships were found between academic major and training, preparation or ability. Since the measures of leadership are all self-report, they must be treated with some caution, but nonetheless, no relationships emerged.

DISCUSSION

It was encouraging to discover that students felt their training influenced their levels of preparation to become leaders, and that they believed classroom instruction was useful. It was also interesting that academic major had no relationship to students' perceptions of their leadership preparation or leadership abilities. The stereotype that science majors have narrow academic experiences and, therefore, tend to be less sensitive to interpersonal concerns was not supported.

Of particular interest were the training issues addressed. It appears that each of the commissioning sources are preparing new officers to some extent to assume their leadership responsibilities. The bulk of this training seems to impact managerial skills rather than the more interpersonal leadership skills. LMET at SWOS furthers this leadership preparation with a positive impact on the interpersonal dimensions. If the Navy's goal is to provide the most efficient training pipeline, it appears that the commissioning sources would do well to concentrate formal leadership training in the managerial skills, leaving the interpersonal skills to LMET. This would be an improvement over providing minimal training in all areas at both schools. It also seems that an interpersonal perspective in managerial training would be worthwhile (e.g., planning work and setting goals for subordinates). Further, while the length and method of the training experiences differ at the various commissioning sources, they would do well to agree on a standard set of leadership issues to be addressed in leadership training and do them as well as possible in the time frame provided.

While the findings from this study are based on self-report they are suggestive of issues worthy of future pursuit. This work will be followed up with input from SWOS instructors as to students' leadership abilities, an evaluation of the shortened LMET curriculum and optimally, a one-year follow up of individual performance in the fleet.

REFERENCES

Butler, M. C., Bruni, J. R., Hartman, E. A., & Hilton, T. F. (1984). Manifest needs among health care professionals: Dimensionality and reliability. Paper presented at the Annual Meeting of the APA, Toronto.

Byron, J. L. (1985). Warriors. Proceedings, June 1985, pp. 64-68.

Sarkesian, S. C. (1985). Leadership and management revisited. The Bureaucrat. Spring, 1985, pp. 20-24.

Steers, R. M., & Braunstein, D. N. (1976). A behaviorally-based measure of manifest needs in work settings. Journal of Vocational Behavior, 9, 251-266.

Taylor, R. L., & Rosenbach, W. E. (1984). Military leadership: In pursuit of excellence. Boulder, Colorado: Westview Press.

BEST COPY AVAILABLE

Leadership Development of USAF Aircraft Maintenance Officers

Captain Michael A. Morabito
437th Organizational Maintenance Squadron, Charleston AFB, SC

Captain Benjamin L. Dilla
Air Force Institute of Technology, Wright-Patterson AFB, OH

A sample of 320 Air Force aircraft maintenance officers (AMOs) were surveyed using the updated version of Yukl's Managerial Behavior Survey (MBS), to measure leader behavior of the AMO's superior officer, and other scales focusing on the AMO's perception of his/her own leadership development. Specific development methods used by AMOs and the perceived importance of each were explored. Furthermore, suggestions were collected on ways to improve development methods available to them in the Air Force. The leadership development activities were correlated with the superior's leader behavior and with demographic and organizational variables. The personal factors of age and rank were found to be associated with leadership development. Participation in 8 of 19 leadership activities correlated significantly with the degree of importance placed on the activities. Analysis of the MBS results indicated certain categories of the superior's leader behavior were significantly associated with the perceived leadership development of the AMO.

Introduction

Leadership is a constant concern in the military environment, yet very little is known about how leaders are developed. The overall objective of this research was to identify methods of leadership development used by Air Force officers, specifically junior aircraft maintenance officers (AMOs), with special interest in the impact of the immediate superior's leader behavior on the junior officer's development.

For the purposes of this research, leadership is defined as a dynamic, goal-directed process of influence between leader and follower, including the interaction of each with the situation (Yukl, 1981). Leadership development is broadly defined as any method or activity used by individuals to enhance their personal ability to influence subordinates toward goal accomplishment.

Leadership Theory and Research

The subject of leadership has been addressed by many scholars from ancient times to the present. One of the major approaches which has been key in understanding the concept is the leader behavior approach. Research at Ohio State University resulted in identification of the classic dimensions of consideration and initiation of structure (Halpin and Winer, 1957), while efforts by the University of Michigan Survey Research Center produced the distinction of job-centered and employee-centered leadership (Likert, 1961) and, later, the four categories of support, interaction facilitation, goal emphasis, and work facilitation (Bowers & Seashore, 1966). These categorizations, although important in developing an understanding of leadership and still used today in the context of some of the situational theories, are not without their problems.

One of the key issues has been the recognition that ". . . these broadly defined categories provide too general and simplistic a picture of leadership. They fail to capture the great diversity of behavior required by most kinds of managers and administrators" (Yukl, 1981, p. 120). This realization led Yukl and his colleagues to develop a more comprehensive categorization, which currently includes 13 dimensions. According to Yukl, "The advantage of the new taxonomy is that it has a larger number of more specific behavior categories than earlier ones, and it includes most behaviors found to be important in leadership research" (Yukl, 1981, p. 128). Because of these advantages, Yukl's Managerial Behavior Survey (MBS) was selected for the description of superior officers' leader behavior in this research.

Military Leadership Development

Studies of the process of leadership development in the military services has focused primarily on professional military education (PME) or other formal programs. The Air Force has essentially come full circle in their philosophy of educating leaders in the past 12-15 years. Reports in the 1970s expressed concern by senior officers on the need for development of technical and management skills (Dobias, 1974; Robinson, 1974) and suggested better pre-commissioning training and even on-the-job training. Then came concern in the late 1970s, still present today, that the Air Force had gone too far in the technical training of young officers, making them more occupational than truly professional (Gosnell, 1980). Another recent Air Force study of particular interest emphasized the important role played by commanders in serving as leadership models for their subordinates (Benton, 1981). This important concept of mentoring has been much neglected and even resisted by the Air Force with its emphasis on formal programs of education and training.

Research within the U.S. Army reinforces this emphasis on the supervisor-subordinate relationship. In analyzing problems in junior officer development, Wellins and others (1980) noted that a key issue was that "the senior officer may not take time to supervise, guide, and correct the performance of the new lieutenant" (p. 5). Another important study conducted within the Army concluded that the most successful leader training results from experiential processes ("learning by doing") rather than analytical or procedural processes (Shriver et al., 1980).

Perhaps the most revealing observation from a review of the literature on leadership theory and leadership development is that there has been little if any research to identify methods which present or prospective leaders actually use for leadership development. To identify these methods and their relative usefulness for a specific group of leaders was the focus of this research.

Methodology

Sample

The original population of interest was all U.S. Air Force aircraft maintenance officers serving in the grade of lieutenant or captain. For practical considerations, this was narrowed to those serving in the continental United States (CONUS) in the three largest major commands--Military Airlift Command (MAC), Strategic Air Command (SAC), and Tactical Air Command (TAC). Generalization of the findings of this research should be limited to these specific population parameters. According to official sources, the population size was 730 officers. Air Force officials authorized

surveys for a random sample of 320 individuals to allow for a 95 percent confidence level.

Procedure

Surveys were mailed to officers at their duty addresses, accompanied by a cover letter signed by the Dean of the School of Systems and Logistics, Air Force Institute of Technology (AFIT/LS). Participation was voluntary, and respondents were assured of anonymity. Respondents were asked to complete and return an optical scanning sheet for the standard survey items. Space for comments and suggestions was provided at several points in the questionnaire booklet. Respondents were instructed to return all materials in a postage-paid return envelope.

Measures

The MBS, provided by Dr. Gary A. Yukl of the Business School, State University of New York at Albany, was used to assess subordinate perceptions of their superior officer's leader behavior. This instrument measures the frequency of 130 leader behaviors (ten items in each of thirteen categories).

The rest of the survey was composed of items written especially for this research effort. It included: (a) standard demographic information, including sex, age, source of commission, rank, major command, prior service, and organization/level (7 items); (b) perceived extent of leadership development (4 items); (c) immediate superior's leadership effectiveness (1 item); (d) perceived importance of leadership development activities (18 items); and (e) extent of involvement in leadership development activities (15 items). Items on PME completed by different methods were grouped together in section e, producing a smaller number of items than in section d.

Results

Respondent Profile

From the random sample of 320 officers, 185 usable surveys were returned, for a return rate of 57.8%. This was considered reasonably good since the survey package contained 27 pages and 190 items. The sample was predominantly male (84%) with the median age category being 30-34 years. Most received their commission from Officer Training School (58%) vice ROTC (37%) or USAFA (4%). Consistent with this pattern, the majority had prior enlisted service (56%). Almost half (48%) were captains, with about an equal split among the rest between second lieutenant (26%) and first lieutenant (25%). Half were assigned to TAC (51%) with the rest divided almost equally between SAC (25%) and MAC (24%). The largest number were assigned to Organizational Maintenance or Aircraft Generation Squadrons (37%); almost two thirds (65%) were assigned to the unit level versus the DCM staff or a higher headquarters.

Leadership Development Activities

Respondents were asked to rate the perceived importance of leadership development activities on a five-point scale ranging from "not important" (1) to "extremely important" (5). They were later asked to indicate their completion of certain activities (e.g., Squadron Officer School) or extent of

involvement in other ongoing endeavors (e.g., number of hours per week spent in personal study of leadership). Results are summarized in Table 1.

Table 1
Leadership Development Activities

ACTIVITY	MEAN IMPORTANCE RATING	INVOLVEMENT (Average or Percent)
Leadership of NCOs	4.21	6.60 Times/Week
Observation of Superiors	4.15	4.19 Times/Week
TDY Deployments	4.15	3.83 Weeks/Year
Leadership of Airmen	4.08	6.65 Times/Week
Leadership of Peers	3.91	4.07 Times/Week
ACSC in Residence	3.86	0.0% Have Attended
Other Leadership Activities	3.39	2.23 Hours/Week
ACSC by Seminar	3.12	8.6% Have Completed
Graduate Degree	3.07	29.7% Have Completed
Personal Leadership Study	2.90	1.73 Hours/Week
ACSC by Correspondence	2.78	5.4% Have Completed
Sports Leadership	2.73	1.27 Hours/Week
Other AF-Related Activities	2.73	0.67 Hours/Week
Community Leadership	2.71	1.02 Hours/Week
Professional Org. Leadership	2.68	1.02 Hours/Week
Church Leadership	2.66	0.69 Hours/Week
SOS in Residence	2.61	34.6% Have Attended
LPDP	2.53	15.1% Have Completed
SOS by Correspondence	2.29	34.0% Have Completed

The most important activities to the AMOs were their working experiences with NCOs, airmen, superior officers and peers, and on TDY deployments. The least important in their eyes were the formal programs which are typically emphasized in official circles. These included the Lieutenants Professional Development Program (LPDP) and PME, especially by correspondence.

Chi-square tests were employed to determine if any significant relationships existed between the rated importance of leadership development activities and the extent of involvement in those activities. Eight of the nineteen leadership development activities were found to have a statistically dependent relationship at the .05 level of significance. They were: leadership of NCOs, TDY deployments, other leadership activities, graduate degree, personal leadership study, other Air Force related activities, professional organization leadership, and church leadership. These activities spanned the spectrum of involvement ratings and level of involvement, so no pattern was obvious. The findings could be interpreted to say that the officers perform a given activity because they perceive its importance in developing their leadership ability or that the activity is rated as being important simply because the officers are involved in it.

Managerial Behavior Survey Results

Before using the MBS results, internal consistency reliabilities were computed for each of the thirteen scales; all were .86 or greater. Also, means and standard deviations of the scale variables (formed by summing the ten item scores for each scale) were computed to provide a picture of the average leader behavior of the superior officers. Results are summarized in Table 2. Of interest in the present context was the fact that "developing" behavior was one of the lowest rated categories.

Table 2
Managerial Behavior Survey

Leader Behavior Scale (10 items each)	Cronbach's Alpha	Mean	Standard Deviation
Supporting	.93	2.86	.726
Problem Solving & Crisis Mgt.	.90	2.78	.632
Interfacing	.92	2.68	.669
Representing	.92	2.67	.700
Informing	.89	2.64	.684
Consulting & Delegating	.91	2.58	.706
Monitoring Operations	.86	2.51	.656
Harmonizing & Team-Building	.93	2.49	.747
Recognizing & Rewarding	.92	2.48	.770
Planning & Organizing	.91	2.40	.755
Motivating Task Commitment	.89	2.34	.731
Developing	.92	2.27	.797
Clarifying Roles & Objectives	.90	2.21	.765

Next, a set of correlations were computed between the AMO's perceived leadership development and the reported leader behavior of the superior officer. The hypothesis was that some types of leader behavior (e.g., developing, consulting and delegating, recognizing and rewarding) would be positively related to the subordinate's leadership development. In fact, all of the correlations between the MBS scales and self-rated leadership development were negative. Nine of the thirteen were statistically significant at the .05 level; however, the largest negative correlation was only -.19 (for "supporting" behavior).

Personal Determinants of Leadership Development

The final major analysis was accomplished to determine if there were any relationships between the demographic items and leadership development by the AMO. A one-way ANOVA was used for each of the categorical variables with perceived leadership development as the dependent variable. The ANOVA for rank resulted in significant differences between group means ($p<.02$), while there was a marginally significant difference for prior enlisted service ($p<.09$). In both cases, greater experience (as indicated by higher rank and prior service) was associated with greater leadership development.

Discussion

Our major conclusion is that experience is the best form of leadership development. The AMOs in this study placed the highest importance on their working relationships with airmen, NCOs, and peers, their observation of superior officers, and their experiences on TDY deployments. Furthermore, these were the activities in which they reported the greatest involvement. These results were also supported by demographic data which indicated greater perceived leadership development for those having higher rank and those having prior enlisted service. It would appear that the old adage, "there's no substitute for experience", is indeed true when it comes to leadership.

A striking result on the opposite end of the spectrum was the comparatively low ratings given to formal development programs such as LPDP

and PME, especially for the correspondence versions. These trends were supported by many of the written comments which emphasized the need for experiential training.

Even other varieties of leadership experience, e.g. in the community, churches and professional organizations, received relatively low ratings versus those categories which were specific to the job. These trends would seem to support the now well-accepted situational aspect of leadership. Experience, to be most effective, should be specific to the job environment.

A surprising outcome was the negative correlations observed between the superior's leadership behavior and leadership development of the subordinate. It was expected that at least some forms of leader behavior would enhance the AMO's leadership development. The negative relationship observed could be explained in at least two ways. They could indicate that if a junior officer has developed well as a leader, then the superior does less leading and allows the junior officer to lead. Another possible explanation is that the more the AMOs feel they have developed as leader, the more critical they are of the superior's leader behavior.

This research focused on rather global issues of leadership development, specifically in the aircraft maintenance career field. Much has been learned; however, many questions remain. Future research should further examine the relative worth of different methods of leadership development; such trends should be examined in various settings with different groups of leaders. A criterion measure more objective than self-rated leadership development would be particularly useful in follow-on work.

References

Benton, J.D. (1981). Promoting leadership in the Air Force's management environment. Unpublished research report, No. 0230-81, Air Command and Staff College, Maxwell AFB, AL.

Bowers, D.G., & Seashore, S.E. (1966). Predicting organizational effectiveness with a four-factor theory of leadership. Administrative Science Quarterly, 11, 238-263.

Dobias, L.J. (1974). An analysis of management development in the Air Force. Unpublished research report, No. 0785-74, Air Command and Staff College, Maxwell AFB, AL.

Gosnell, W.L. (1980). The Air Force is making occupationalists of its junior officers. Unpublished research report, No. MS071-80, Air War College, Maxwell AFB, AL.

Halpin, A.W., & Winer, B.J. (1957). A factorial study of the leader behavior descriptions. In Stogdill, R.M., & Coons, A.E. (Eds.), Leader behavior: Its description and measurement (pp. 39-51). Columbus: Bureau of Business Research, Ohio State University.

Likert, R. (1961). New patterns of management. New York: McGraw-Hill.

Robinson, G.D. (1974). Schooling the middle manager. Unpublished research report, No. 5406, Air War College, Maxwell AFB, AL.

Shriver, E.L., and others. (1980). Development of a leader training model and system. Unpublished research report, AD-A082 730, Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Wellins, R.S., and others. (1980). Analysis of junior officer training needs. Unpublished research report, AD-A096 034, Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Yukl, G.A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

IMPLEMENTING TECHNOLOGY IN A MILITARY ENVIRONMENT

William L. Maloy and Nancy N. Perry

Chief of Naval Education and Training
Pensacola, Florida

INTRODUCTION

The Chief of Naval Education and Training established the Training Technology Implementation (TTI) Office on his staff to transition promising technologies into Navy training settings systematically. The charter of this organization includes placing specific technologies and techniques in the training environment, and, more importantly, developing successful strategies and overall architecture for implementation to guide future efforts.

This paper briefly describes three ongoing projects in order to form the context for examining several lessons learned by the TTI office. Out of these lessons surface broad policy issues which, if dealt with, promise more orderly implementation of training technologies and techniques in the future.

IMPLEMENTATION PROJECTS

Automated Maneuvering Training Board System

The first TTI initiative grew out of the Defense Science Board's Summer Study of 1982 pronouncement that not enough technology was finding its way into the military instructional environment. The Chief of Naval Research agreed to provide investment capital for this endeavor. The Chief of Naval Education and Training was assigned the task of designing the implementation architecture and directing the implementation process. Representatives from the Chief of Naval Operations, the Chief of Naval Education and Training, the Office of Naval Research, and the Navy laboratories identified several of the most promising Research and Development (R&D) projects as candidates for this implementation project.

Meanwhile, researchers at the Navy Personnel Research and Development Center were developing a new method for guiding the conceptualization of ship maneuvering interactions: the Automated Maneuvering Board Training System (AMBTS). Because of the inherent difficulty involved in mastering maneuvering board skills, the implementation of AMBTS at the Operations Specialist (OS) "A" School in Dam Neck, Virginia was selected for the initial demonstration project. Implementation architecture included training effectiveness and efficiency analyses, life-cycle costing, and acceptability.

BEST COPY AVAILABLE

Performance of students using AMBTS is being compared to that of students receiving the traditional instruction. Additionally, a sample of students will be followed into the fleet to determine if performance differences persist in the job environment.

AMBTS in Pierside and Reserve Training

The second initiative grew out of the technology of the first, but focuses upon the pierside and reserve community components of the training continuum. The issues here are not the effectiveness and efficiency of initial skill training but rather maintenance of skills, recertification of individuals returning to sea after protracted shore assignments, and qualification of teams whose players must learn to coordinate their individual skills to maximize team efforts. The fact that the Reserves must accomplish this training at remote sites exacerbates these problems for them.

AMBTS is the medium for accomplishing the training; however, a method for assessing individual performance and prescribing follow-on or refresher training must be added to the capability of the basic program. Through this application of the technology, we hope to learn how to apply what is known about the retention of learning. This will help us design acceptability strategies which can be applied in the fleet and the reserve community where there is a good deal of skepticism about technology's value as an instructional tool.

Training Continuum Strategies

The third initiative is in an entirely different training arena, but builds upon the training concepts of the second initiative. In this case, the implementation model focuses not upon the learning and maintaining of technical skills but on the educational requirements of the officer accession continuum. We have selected an instructional continuum that begins with a Navy school preparing potential NROTC students for the college environment, to the NROTC itself, to the follow-on warfare schools, and, ultimately, to the fleet.

The process for making this continuum operate more effectively includes: (1) assessing individual capabilities at each school level against required developmental and professional skills; (2) intervening at each point along the way to close deficiency gaps through the application of computer-based modules and other good support practices (including peer tutoring); and, (3) providing follow-on assessment and deficiency prescriptions to the next schoolhouse which can then continue the remediation/development process.

Through this interaction among those responsible for training along the continuum, we seek to add a "feedforward" as well as a feedback process that establishes an

accountability for value added at each follow-on school. If we are successful at packaging these techniques, we will have in place the basic ingredients of a true training continuum which should become the basis for all Navy training in the future. Perhaps we can then, at last, reduce the impact that manpower and personnel policies have upon training performance and concentrate upon helping people learn what they need to know -- without wasting time on blame.

TTI has other initiatives underway. These include the identification of applications for smoke generators in Reserve training, the employment of artificially generated signals in gram analysis training in Navy settings, and the development and implementation of a Navy-related functional skills math and reading curriculum.

The purpose of all these early initiatives is to package existing technologies and instructional materials into meaningful instructional delivery systems. These systems are tailored to the unique instructional environments along the continuum and enjoy the credibility and acceptability of schoolhouse instructors and fleet supervisors. As we continue to improve this packaging process and couple it with such delivery concepts as requalification, we come closer to exploiting our ability to intervene in the capability of people to perform. We can then allocate our resources on continually assessing, building upon, and sustaining skill development as the cornerstone for fleet readiness.

IMPLEMENTATION LESSONS

During planning, initiating, and executing several implementation projects, TTI encountered recurring issues and problems. By examining and trying to solve them, we uncovered lessons that should be heeded during systematic implementation. These lessons are consistently found and persistent enough that even the obvious ones should be discussed and all should be incorporated into overall implementation models. Attending to them leads to nothing more than good implementation practices.

1. The difficulty of implementing technology is continually underestimated.

This is a self-evident truth, but one that is generally ignored. Researchers and operators in both the military and civilian communities continually underestimate the difficulty of implementing technology. Researchers and operators have different goals and interests; therefore, just recognizing that the process is not likely to solve the problem.

Perhaps this is why we are seeing a new breed of professional; and the lesson is we must encourage the

development of these people -- people who can stand, however shakily, with one foot in the the research laboratory and the other in the operational environment. These specialists, employing these emerging implementation skills, can bring the transitioning process to a level of sophistication which will build a climate of operator acceptance and ownership. Maintaining this climate throughout dissemination, will best contribute to the sustainability of new training technology and techniques in the classroom.

2. The capability of technology to improve training should not be exaggerated.

This second lesson highlights a principle which has long been discussed with regard to innovation but is often overlooked in practice. The training equipment that can be uncovered in military salvage depots is testimony to our willingness to embrace new technologies as panaceas and our subsequent disillusionment with them.

Similarly, in our zeal to adhere to the letter of our instructional systems design law we have too often confused familiarity with mastery. We have also overlooked the impact of forgetting on initial performance in an operational setting. It is now apparent that exaggerated claims of improved performance, often accompanied by poorly utilized computer technology, helped to undermine the credibility of those endeavors with operators who knew better. This frequently resulted in animosity toward the ISD process and the techniques and technologies it promoted.

At the beginning of every implementation endeavor we are well advised to realistically formulate promises and use technology in supporting roles and as tools where appropriate. Our initiatives at TTI incorporate this advice.

3. Goals determining hardware and software uses must be carefully defined.

It is important to have goals that determine hardware and software uses. What we are learning, however, is that we must use great care not to structure our goals too narrowly around near term cost and training benefits as we perceive them which may, in turn, restrict our our ultimate flexibility.

As we get better at recognizing technology linkages, there will be numerous opportunities to build one system upon another. In this way we can expand the applications for individual team and crew training uses as well as leapfrog those applications from the schoolhouse to shipboard environments. However, it will be difficult to fully exploit these possibilities if we have been overly restrictive in our early-on goal setting and front-end

analysis processes. We must improve our abilities to perform trade-off analysis and take risks.

4. Technical skills must be acquired in order to implement technologies successfully.

Trainers and implementers may have an edge over others when it comes to analyzing the capabilities of technologies to improve training, but specific hardware decisions are often out of our realm of expertise. We need people who know about hardware configurations -- people who can provide guidance in ordering equipment and helping us with hook-up, operating, and maintenance problems.

The role of these people includes the development of quality controls in the acquisition, use, and maintenance of hardware and materials, and monitoring vendor performance. There is no reason to tolerate the shoddy performance that many of us have experienced in training systems technology. We should use these people to develop stringent, enforceable quality controls across the board.

POLICY ISSUES

Our early attempts to move technology from research to implementation brought two policy issues into bold relief. If our initiatives are to be optimally successful we must deal with both the issue of a funding continuum to accompany the R&D continuum and with procurement issues.

Funding Continuum Issue

As an R&D product moves from exploratory research to engineering development, to prototype demonstration, to pilot testing, and, finally, to full-scale implementation, a systematic funding continuum to assure this orderly progress is essential but currently lacking. To establish this funding continuum we must begin by identifying clearly stated, broadly based umbrella-type research categories under which specific projects can be grouped. These categories should be derived primarily from senior officers responsible for training who should review the categories periodically in order to renew their commitment or change them as a result of changing military priorities.

Specific R&D projects will fall under these umbrella categories where they will be prioritized and appear as a totally funded package through the 6.4 funding level. This portion of the continuum must then be coupled with follow-on operating dollars, initially in the form of wedges. As the budget year comes closer, these wedges will give way to the more specific dissemination and life cycle requirements emerging from the 6.4 accomplishments.

Without this kind of funding continuity we will never fully realize the potential that research-based instructional techniques and technologies hold for improved performance.

Procurement Issues

Software and hardware procurement and maintenance issues can no longer tolerate policy neutrality. In the early stages of transitioning R&D from the laboratory to the classroom, trainers must be formally involved in the procurement process. There are too many training issues involved in purchasing decisions (and many quality control initiatives) to leave technology skilled trainers out of the process.

Further, we must weigh the advantages and disadvantages of the centralized/decentralized issue in upgrading and maintaining software and courseware. To date, each implementation project has found its own, not always satisfactory, solution to this problem. We must consolidate the knowledge gained by these projects and prepare a standardized, supportable policy on this issue.

Finally, policy matters associated with life cycle costing greatly concerns those responsible for training in the classroom and at pierside. How these policies evolve is closely tied to the credibility and acceptability technology applications will enjoy among our military colleagues.

CONCLUSION

We in Navy training have embarked upon implementation in a formal way. We believe this to be good, sensible, and even wise. However, we suspect that our initiatives, our experience, and our policy concerns differ little from others'.

The task before us, then, is to learn how to learn from each other; to share what we know; and to build on each other's experiences. How well we do these things will determine how well we fulfill our goals in the struggle to be more effective in a world of scarce resources where so much depends on human performance.

Research Needs Assessment and Technology Transfer in USAREUR

Karol Girdler and LTC Ford McLain
U.S. Army Research Institute

In October 1984, the Army Research Institute (ARI) Field Unit in the U.S. Army Europe (USAREUR) ended its programmatic research activities, and it was redesignated as a Scientific Coordination Office (SCO). The primary mission of the organization became research needs assessment and technology transfer. Technology means scientific knowledge in the form of hardware, processes, methods or ideas. Here, technology transfer refers to the application of technology produced by ARI researchers for a new user and in some cases, putting old results to new use. Since one-third of the U.S. Army is stationed in Europe and numerous ARI technologies are developed for use by the active Army in the field, it is essential that products be transferred and needs be considered within this setting. While identifying research needs and promoting technology transfer, we have identified gaps in what should be a comprehensive research and development (R&D) management process. The purpose of this paper is to clarify the need for an R&D management approach which is accountable for research including needs assessment and product utilization with feedback from the user.

Highlighted here are our conclusions developed as we searched for a model against which we could test our experiences in needs assessment and with which to guide and judge our effectiveness in disseminating knowledge and facilitating implementation of research products in an operational setting. In searching for a model, we were, in part, seeking to define the role of and the support system needed for those who carry out these functions in the military setting.

Dissemination of research results, utilization and the relationship to needs assessment are not new ideas. A great deal has been written on ways to disseminate knowledge and get it used, and about the critical variables, barriers and gateways to be encountered. Glaser (1983) compiled an extensive review of research on the topics and reviewed several models. The bibliography alone is 162 pages. The demand for better needs assessment analysis and product utilization in the military, in particular, have been addressed by military psychologists such as Freda (1980) and Shields (1977). For example, Shields found that some technology is under-utilized or totally rejected by the intended users because the need for the product did not exist, the product was flawed from a user standpoint, or the researcher or developer did not attend to the complete process of R&D management.

What is new here are our discoveries of what is possible now, and what further support is required as we perform these functions in an operational environment. Also unique is that we may be the only unit in the R&D community with these functions as the main mission of the organization.

This paper is an unofficial document intended for limited distribution to obtain comments. The views, opinions, and/or findings contained in this document are those of the authors and should not be construed as the official position of ARI or as an official Department of the Army position, policy, or decision.

We have speculated and forayed into our environment with a number of simultaneous actions in accomplishing our mission. We have had successes, and at all times we have come back to the central question of how our activities do fit or should fit into a system of R&D management and a model of knowledge utilization. Our aim is to do more than "muddle through" with occasional immediate successes; it is to influence the further institutionalization of our activities and provide wider support for such action throughout the military and the R&D community. Within the military, centralized management of R&D may be too big to handle; but, a change in the sequencing of and habitual relationships within the R&D community and between R&D people and users still needs to be considered.

One way to provide institutionalization of R&D management for this purpose is on a case-by-case basis, that is, a case management system for research products. If such longitudinal management existed in a clearer, more concise form, to include greater involvement with the users, we would be in a better position than we are now. However, our experience suggests to us that product case management may not be the most effective procedure. Such a potentially regulation-based system could solidify horizontal communication in organizations or result in paperwork drills. In addition, the concentration on a product-based system could limit the perception and creativity of those responsible. The longitudinal mindset of such a system could lead managers to see R&D management as a process where each product has a definite start and end unconnected to other products and issues. Rather than a product case management process, the ultimate support system for scientists engaged in these activities would recognize research needs assessment and technology transfer as two sides of one coin, the activities occurring in a cycle and occurring simultaneously for many products. The common element to be managed in this setting of multiple activities could be not products, but concepts in relationship to central issues.

We are interested in institutional support for the scientist operating in the role we are defining - one who deals simultaneously with a number of user needs and research products and seeks to bring products to the intended user, as well as others who could benefit from the knowledge. At the same time, information is transmitted about new users and new needs to those scientists producing technology. These needs might be answered with new research or existing products which can be implemented. This in turn may suggest new needs relevant to the product implemented, as well as to other research, so that the full cycle has occurred. The cycle is not performed for one product and one need. The scientist may have several needs or sets of needs in mind for various organizations, or several possible products and be at various stages of the cycle in different settings. R&D management processes could create a support base for performing these activities. The research process which receives the most support and management now is the normal scientific method or process of problem review, hypotheses generation, etc. The role being described here of making broader linkages and connections takes us to areas which are previous to problem review and which follow research project completion, and even follow product utilization.

The continuing questions are: How do we organize support for the scientist in this role? Who in effect is a research manager? How do we

organize the work itself to keep it flexible and avoid bureaucracy which could stifle communication needed to identify new uses for technology? The role is that of a gatekeeper of established channels of information exchange, and of a gateway creator who must ascertain focal points where "gates" can be created between the huge systems of the military and the R&D community. Where are the interfaces between the two systems most permeable and facilitation most likely to allow information to be passed and used? In this age of information, it might require only a few people to manage these aspects of the R&D process given access to the right tools and the proper vantage point or perspective. The individual in this role needs access to the entire R&D community to assist in managing research in this sense. The database or information network to support such work needs to include not only technical information about current products and their development, but information about issues and trends of primary importance to the military and information about new concepts of technology.

A few examples may help show the reader where we see opportunities to organize successful exchanges of information, to facilitate the process, and suggest management needed for a support base. Keep in mind that exchange means information passed both ways. In effect, the scientist involved in this aspect of research has two major groups of "clients:" The R&D community and the military user.

There are a number of ways new research needs are identified. As noted above, need identification is often performed simultaneously with implementation or information dissemination. At our office in particular, one form this multiple operation takes is what we call Technical Advisory Service (TAS). In this function, military users make requests for our assistance based on their perceived operational needs, and their knowledge of our expertise which we have made known in the community.

Our first example concerns TAS being carried out at the Warrior Preparation Center (WPC). The WPC is a joint operation by USAREUR and the United States Air Force in Europe (USAFE) to support command and battle staff training. The center provides training for the operational level of warfare for corps through Army Groups in joint exercises with the Allied Tactical Air Force (ATAF). Our organization had originally approached WPC to possibly assess needs at this new center which might be addressed by ARI research. After the WPC staff members became familiar with the SCO mission, functions, and areas of expertise, they began to identify operational and developmental requirements which could be addressed by the ARI SCO.

In this case, establishing a relationship and facilitating discussions with WPC resulted in benefits to both organizations. We have been able to provide short-term TAS yielding ideas and answers for WPC operational development, specifically in terms of data collection and data analysis. The work was done by one SCO psychologist and by one scientist "loaned" from ARI HQ for an intensive six-week consultation to WPC. By applying our consultation skills at WPC and other organizations, we build credibility for ARI in USAREUR and lay the foundation for a positive reception of future introduction of ARI technology, and we create support for the process of our work. The WPC organization helps us do our job of identifying research needs as they continue a dialogue with our office, in which they describe exercises, issues of measurement, and performance at corps and echelon above

corps level. In the future, we believe we might be able to introduce to WPC results from ongoing ARI research on corps performance and use feedback from WPC as guidance for further research and development. The lesson learned is that consultation skills and establishment of networks or relationships provide the basis for our role.

Our second example concerns training technology being developed by ARI for the Army Training Battle Simulation System (ARTBASS), which is to be implemented as the Army's primary system for training battalion command groups in the command and control of combat operations. ARTBASS uses a computer to simulate unit actions and weapons effects, and to calculate the changing status of personnel, equipment, ammunition and fuel in simulated combat. The system is portable, and the Department of the Army plans to field enough systems to train every maneuver battalion in the active Army and every Reserve component.

In our larger efforts to clarify issues and needs in command and control below corps in USAREUR, we attempted to ascertain who the users for ARTBASS training technology are in USAREUR and how the training technology being developed by ARI might be fielded along with the new system it supports. Previous discussions with military users in the field had led us to identify an important training issue in USAREUR: Supportive training products frequently arrive much later, if ever, than new training systems or new equipment. The need to make connections was clear. However, our experiences in the transfer of training technology for the ARTBASS raised more issues than answers about the R&D process vis-á-vis our role.

Our attempts to make linkages in USAREUR and to have prototype ARI training products for the ARTBASS examined were well-received by the 7th Army Training Command (7ATC) in USAREUR, the primarily responsible organization for implementation of ARTBASS. Likewise, the Army Materiel Command Europe (AMC) encouraged our coordination with them in their role in fielding the system. Our informal information exchange and coordination appeared to be holding up well within USAREUR; however, a subsequent meeting with a representative from the U.S. Program Manager for Training Devices (PM TRADE) office left us with questions about our role. The representative from the agency primarily responsible for overseeing development and initial implementation of ARTBASS was unaware of who ARI was and why coordination in the USAREUR environment was desirable. Coordination for fielding of training technology developed by ARI was not on the list of "things to do" when fielding new devices. Formally, at least to this representative, ARI was not considered as a contributor to the overall project. When we were questioned about how centralized coordination of training technology and new devices was done in the States before fielding to USAREUR, we had to respond that we did not know. The representative's proposed efforts to rectify the lack of coordination were in terms of a centralized, one product or one case fix, rather than a systems approach. Our coordination with the U.S. representative virtually came to a halt.

At this point we are left wondering. We have been told numerous times that new devices and equipment often arrive in USAREUR without programs of instruction or other types of supportive training materials. At the same time, we were often able to point to finished ARI products which could assist in these situations. It fell to the SCO in Europe to pursue the

connections. We have not yet answered the questions. Should the coordination role be assumed by a central R&D management level? Are many decentralized efforts necessary? When and with whom? We continue to build our linkages in Europe and pursue the expansion of linkages in the entire R&D community. It is possible, however, that we will find ourselves outside the system as we create our informal network.

Our third example concerns a product for which our lateral communications have identified potential new users, and for which we are assisting in a test and evaluation process to boost user involvement in the product's development. The success we have in making connections is based largely on the fact that a new type of program has recently been initiated and based at HQ ARI to provide research product management and which provides some formal mechanisms to support a role such as we are developing.

The Joint Service R&D Program sponsored by DOD was initiated to offer end users the capability to participate in the development of prototype technologies. One product from this program under development jointly by the Navy and Army is the Personal Electronic Aid for Maintenance (PEAM). This highly portable, interactive device would replace maintenance manuals and provide an aid to performance of Organizational Maintenance. The initial application is for the Turret Mechanic of the M1 Abrams Tank.

Our involvement with the Joint Service program combined with our awareness of needs and issues in Europe has permitted the SCO to make connections which may expand the use of PEAM to address a related but different need beyond its planned application. We are currently involved in evaluating the feasibility of expanding the portable, step by step electronic guide into versions for use by German maintenance mechanics who do not use the English language manuals well. In addition, other products based on the concept of easy access to procedural knowledge might become involved in addressing the maintenance issue.

We see the PEAM example as an important aspect of the total R&D process. Within the greater R&D community when these types of connections are made, they tend to be as a result of personal efforts and interests. In the case of PEAM, the Joint Service R&D Program sets the stage for a broader view of research and development. When this type of R&D management is combined with the flexibility to share information and communicate about an issue laterally through military operational settings and R&D organizations, good ideas can possibly surface more easily and become reality.

Our final example concerns our activities at 7ATC where informal linkages built through our use of consultant skills has created gateways through which finished ARI products have been delivered to users.

7ATC is responsible for planning, developing, managing, and coordinating training requirements and programs for USAREUR. An initial visit with 7ATC staff members showed that they were generally unaware of ARI, and yet very interested in the type information ARI was capable of providing. After our learning about 7ATC's mandate and operations from staff members and written materials, we briefed down through the 7ATC chain of command, addressing what we identified as their staff concerns, to establish a formally recognized gateway to this organization.

We began our work with one directorate at a time, first establishing needs and delivering information, to establish a "track record" within the organization of delivering what we promised, before going to the next area. We currently play a consulting role, visiting 7ATC on a routine basis, and being called in to address specific problems. This allows for follow-up of previous work, and identification of new needs. It must be recognized that there is only limited organizational memory within 7ATC directorates. When a staff member leaves, our work essentially leaves with him. Likewise, when a new staff member comes on board we must brief him on our mission and provide a package of materials that identifies what has, and can be done to support staff members' job requirements.

There is still no routine delivery of ARI products at 7ATC for several reasons. The 7ATC staff carries a very heavy workload, and does not have the luxury of reading ARI reports, even when these are readily available. We have learned through experience that staff information needs must be clearly defined, and applicable ARI material discussed one-on-one if the information is to be recognized and applied to the 7ATC mission, and eventually in USAREUR. Using this approach, we have been able to identify and address a number of information needs using past research products and current ARI draft research materials. We have also been able to identify future products, both short-term, and long-term, which address 7ATC needs, and have provided these as they were released. Based on research product feedback and new operational needs at 7ATC, the SCO currently identifies needs for incorporation into ongoing ARI research, thus continuing the needs assessment - technology transfer cycle.

In conclusion, we are looking at research needs assessment, information dissemination, and utilization, which comprise a scientific field of continuing development - one in which a number of models have been suggested but none are definitive. We are attempting to create not only gateways to exchange information, but relationships which can be accessed through the gateways and through which the user can become a fuller part of the R&D cycle to influence products both before and after their creation. As we perform the activities, we are constantly examining our actions and assessing how they fit or do not fit smoothly into overall R&D management. We are searching for system solutions to facilitate our roles. If we do not, we will always be treating the military needs on a band-aid basis, and when we, the gatekeepers leave, the gates may close. By defining our role and identifying the types of support and credibility needed institutionally for the role, we hope to get consistently more use from each research dollar on a long-term basis. The payoff is in effective, user-oriented research management.

References

Freda, J. S. (1980). Army Training Technology Transfer: A Systems Model. (Research Report 1241). U.S. Army Research Institute.

Glaser, E., Abelson, H., & Garrison, K. (1983). Putting Knowledge to Use: Facilitating the Diffusion of Knowledge and the Implementation of Planned Change. San Francisco: Jossy-Bass Publishers.

Shields, J. L. (1976, October). "Training Technology Transfer." Proceedings of the 15th Annual Operations Research Symposium. Ft Lee, VA.

The Reduction of Standard Errors of Equipercentile Test Equating through Negative Hypergeometric Presmoothing*

Benjamin A. Fairbank, Jr. Ph.D.
Performance Metrics, Inc.
5825 Callaghan Rd., Suite 225
San Antonio, Texas 78228

I. INTRODUCTION

Test equating is the process of finding which scores on two or more similar tests correspond to the same level of ability in the population of potential examinees. The need for test equating arises as a result of many considerations. It is often valuable to have more than one version or form of a test. When more than one version or form of a test is available, the particular form taken by an examinee should not affect the examinee's expected score.

The replacement of operational tests requires equating when the scores on the new tests are to be used in the same predictive or evaluative equations or in the same manner as were the old scores. Test equating may be carried out in any of a large number of different ways, some of which are of recent origin and are technically sophisticated, and some of which have been in use for several decades (see Holland & Rubin, 1982). This report addresses only equipercentile test equating as applied to two equivalent groups (Angoff, 1971). The terms "reference test" and "experimental test" are used to indicate, respectively, the test whose score metric is to be used for the results of both tests, and the test whose score is to be converted to the units of the other test. For example, if an existing test known as Form K is to be replaced by a similar test known as Form M, Form K would be the reference test and Form M would be the experimental test.

As with any procedures having the goal of estimating population characteristics based on data obtained from a sample, there are always sample-dependent errors present in test equating. If an equipercentile equating were to be done twice with similar samples, the results would differ. The extent of such differences has been estimated by Lord (1982) and their magnitudes appear as the standard errors of equipercentile equating. As with all standard statistical procedures, the size of the expected errors decreases linearly with the square root of the sample size. It is thus operationally impractical to reduce errors beyond a certain amount by increasing sample sizes. For example, decreasing the error to one-fourth the size of the error associated with a given sample size would require using a sample 16 times the size of the original sample. As a consequence, practitioners of equipercentile test equating have looked for other ways to reduce equating errors. They have most frequently used the methods of smoothing.

Smoothing

Two general classes of smoothing methods were used. A third class is made up by combining a smoothing method from the first class with one from the second class. First, presmoothing is defined as the process of smoothing the observed score frequency distributions prior to the equating. Second, postsmoothing is defined as the process of

*This report is excerpted from a technical report by B. Fairbank, Jr. titled "Equipercentile Test Equating: The Effects of Presmoothing and Postsmoothing on the Magnitude of Sample-Dependent Errors," AFHRL-TR-84-64, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, 78235. The research here reported was supported by a contract between the Air Force Human Resources Laboratory and Performance Metrics, Inc. The opinions are those of the author. The full technical report is available from the author.

smoothing the equipercentile points after equating. Third, combined smoothings involve presmoothing and postsmoothing applied consecutively. The common intent of all three smoothing methods is to remove small sample-dependent fluctuations from the nonsmoothed equatings so that the small sample equatings will more nearly approximate the asymptotic equatings, or those which would result from the use of samples so large that the sample-dependent errors approach zero. The extent to which the various methods achieve this common intent is investigated by this research. Seven presmoothing methods, seven postsmoothing methods, and five combined smoothing methods were used as follows:

A. Presmoothing Methods
   1. 3-point moving medians
   2. 5-point moving medians
   3. 3-point moving weighted averages
   4. 5-point moving weighted averages
   5. 5-point moving weighted averages with root transformation
   6. 4253H Twice
   7. negative hypergeometric

B. Postsmoothing Methods
   1. linear regression
   2. quadratic regression
   3. cubic regression
   4. orthogonal regression
   5. logistic ogive
   6. cubic splines
   7. 5-point moving weighted averages

C. Combined Smoothers
   1. negative hypergeometric + orthogonal regression
   2. negative hypergeometric + quadratic regression
   3. negative hypergeometric + 5-point moving weighted averages
   4. 3-point moving weighted averages + 5-point moving weighted averages
   5. negative hypergeometric + cubic splines

The final presmoothing method (see Keats & Lord, 1962; also Lord & Novick, 1968, pp. 515-520) is one devised explicitly for smoothing or fitting frequency distributions of test scores. The distribution is the negative hypergeometric, whose appropriateness is derived from a binomial error model of test scores. The model assumes several technical conditions, one of which is equivalent to the assumption that all of the items on the test whose score distribution is being fit are equally difficult. That condition is known to be false in the case of most tests, but the fit of the negative hypergeometric is still good enough to make it promising for further study (Keats & Lord, 1962).

Objectives

The aim of the present effort was to evaluate the effects of various different methods of presmoothing, postsmoothing, and combined smoothings on the accuracy of test equating. The study was exploratory in nature, designed to determine which methods hold the most promise for operational use.

II. METHODS

General Plan

The plan underlying this investigation was to use three different approaches to determine the effectiveness of each of 14 unitary smoothing methods and five combined smoothing methods. The first approach used simulated tests and examinees; the second

and third used data from tests administered to examinees under operational conditions. The advantage of simulated tests and examinees is that all quantitative aspects of the tests and examinees are completely specified, and it is possible to know in advance the results of theoretically errorless equatings or those equatings which are unaffected by sample-dependent errors. Operational data, of course, have the advantage that they are obtained under conditions typical of the ones under which smoothing methods would be used. The data are not based on an ideal model, as are the data from simulations; rather, they contain all of the departures from theory that may be found in operational test settings.

The first of the three methods of evaluation involved comparing each of the smoothed equatings with a known errorless equating. The known errorless equating was based on a method that yielded results typical of an equating using an infinitely large sample. The method requires deriving a distribution of expected score frequencies, the distribution being that which would result from administering the test to a sample so large that the observed proportions at each score were observed essentially without error. The results of the simulated test administrations were used for that method. The second method was a similar comparison of sample and criterion equatings, but in place of data based on simulations and on an errorless equating, the comparison used operationally obtained data and an equating based on an unusually large sample size. The third method was to use the statistical jackknife (Mosteller & Tukey, 1977) to estimate the size of standard errors of smoothed and unsmoothed equatings using operationally obtained data and simulated data. Those errors were also compared to standard errors computed by means of the formula given by Lord (1982).

Each of the six simulated examinations was "taken" by 100 groups of 2,000 simulees. Either of two methods was used to administer a test in simulation. The first method is similar to that given by Ree, (1980). The second method involved taking a sample of 2,000 observations at random from the Expected Observed Score Distribution (EOSD, see below) for a test. Score distributions were tabulated for each simulated administration. For each test length, 100 equipercentile equatings and smoothings were then performed using the methods described below.

The method used to establish the criterion equatings for the simulations used in the present study is based on the EOSD for each test. An algorithm developed by Lord and Wingersky (1983) was used to prepare distributions of expected observed scores for each of the six simulated tests. In an EOSD, each score has associated with it a proportion of examinees, not a frequency. The distributions model the result of administering the test to an infinitely large number of examinees and observing the relative frequency of each score. By comparing the small sample equatings (N=2,000) with those that result from an "infinite" sample (i.e., those based on the EOSD), the extent of improvement resulting from smoothing is directly observable. The criterion equatings, then, are the unsmoothed equipercentile equatings which result from using the EOSDs in the unsmoothed equipercentile method.

Operational Data

The operational data that were used were taken from a set of test scores for very large sample sizes (approximately 100,000 examinees) for three roughly parallel forms of each of several subtests. Among those subtests were two forms of one test of length of 25 items, and two forms of another test of length of 20 items. In addition to the frequency distributions of test scores for all examinees, there were available 100 samples of 2,000 scores for each of the four subtests. The samples were drawn at random without replacement from the larger samples of 100,000 examinees.

For the operational data, criterion equatings were established by using the full sample of 100,000 examinees. As with the simulated data, the criterion equatings were unsmoothed equipercentile equatings. As with the simulated data, 100 reduced sample equatings were made for each of the test pairs, both without smoothing and with each of the 19 smoothing methods.

Equatings

All test equatings were performed using the equipercentile method described by Lindsay and Prichard (1974). For the unsmoothed equatings and the equatings to which only postsmoothing was to be applied, the raw frequency files were equated. When the equatings involved presmoothing, the smoothed frequency estimates were equated. Following the equatings and smoothings, each test or simulated test had associated with it a criterion equating, an unsmoothed equating, and 19 smoothed equatings, one for each of the smoothing methods used.

Analysis of Equating Results

Each of the five tests, three simulated and two operational, had associated with it one criterion equating, 100 unsmoothed equatings based on sample sizes of 2,000 (called the "small sample"), and 100 sets of 19 smoothed equatings based on the same samples. The question of interest is the effect of the smoothings on the accuracy of the equatings. A deviation is a difference between an equated score obtained with a small sample and an equated score based on a criterion equating. At each observed (i.e., integer) score on the experimental test, the corresponding score on the reference test was found using the criterion equating. The equated scores were found as decimal fractions not rounded to the nearest integer. The score corresponding to the same experimental test score was then found for the unsmoothed small sample equating and for each of the 19 smoothed equatings. The differences between the equated score based on the criterion equating and the equated score based on the small sample equatings were found for each possible score on the experimental test, for the unsmoothed and for the smoothed equatings, for all 100 replications. These differences, or deviations, were the raw data used for evaluating the smoothings. For each of the 100 small sample equatings, the deviations at each score were combined across equatings to give a general measure of deviation at each score. Three such deviation measures were computed.

The first measure is the Root Mean Square Deviation (RMSD), found by taking the square root of the sum of the squares of the deviations across all 100 samples. The second measure is the Average Absolute Deviation (AAD), or simply the mean of the absolute value of the deviations computed across all samples. The third measure is the average of the signed values of the deviations (ASD), found by taking the mean of the deviations across all 100 replications. ASD differs from AAD in that the absolute values are not found before the mean is computed. ASD is sometimes called "bias," or "statistical bias," but in the context of testing the term "bias" denotes other phenomena and so is less appropriate than "ASD."

III. RESULTS AND DISCUSSION

It was found that with some smoothing methods, especially the presmoothing methods, smoothing resulted in large increases in the deviation measures for very low test scores. In some cases the increases were so great that graphing them required such a large rescaling of the figures that the more important deviations in the middle ranges of the test could not be represented. These large induced deviations are seen as being of little interest because they occurred at score values which were lower than the guessing level on a test, and so were not associated with meaningful measures of ability. In order to show the more relevant deviations effectively, the figures do not present information on the levels of RMSD, AAD, or ASD at test scores below the guessing level for each test.

Of the 14 smoothing methods, negative hypergeometric smoothing was uniformly most effective in reducing root mean square error. The results of that method are shown in Table 1.

Table 1. Summary of the Averaged Effects of Presmoothing by the Method of Negative Hypergeometric

Test Length	Proportion of Mean Deviations RMSD	AAD	ASD
		Simulated Tests	
15	.891	.903	2.919
30	.865	.867	3.596
50	.852	.861	3.453
		Operational Tests	
20	.905	.908	2.008
25	.966	.989	7.479
Mean	.896	.906	3.891

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing. Averages taken over all samples at all scores above chance level.

To evaluate the effects of smoothing, particularly its effects on deviations, it is helpful to consider such deviations within the context of the accuracy of ability or achievement tests more generally. The standard errors of equating discussed in this report are not the only measurement errors which arise in the testing process. There are also standard errors of measurement that are intrinsic to any test which is not perfectly reliable. The following formula (Allen & Yen, 1979) relates reliability (R), standard error of measurement (SE), and test score standard deviation (SD).

$$SE = SD * \sqrt{1 - R} \quad .$$

Thus, the standard error of measurement for the experimental test of length 15, based on a reliability (KR-20) estimate of .80 and a standard deviation of 3.28, is 1.47. Similarly, the standard error of measurement for the experimental test of length 30 is 2.20, and that for the experimental test of length 50 is 2.74. The corresponding average standard errors of equating, as estimated by Lord's formula, are .15, .30, and .51. Thus the standard error of equating ranges from approximately only 10 to 20 percent of the standard error of measurement.

The results of smoothing by the negative hypergeometric are the only ones which show consistent improvement in RMSD and AAD as a consequence of smoothing. The effects are particularly impressive with the simulated tests, presumably in part because the criterion equatings for those tests are nearly perfect, not estimated from very large samples. The gains are not uniform across the tests. On the shorter tests at lower scores, the measures of RMSD and AAD actually increase as a consequence of using the negative hypergeometric. The beneficial effects of the negative hypergeomtric do not extend to the measures of ASD. The ASD increases both globally and locally, sometimes quite dramatically. These increases were expected at the lower end of the test, where guessing is a factor, but increases at the upper end were not expected. It must be noted, however, the ASD figures were low initially, so that a tripling of ASD may still denote an acceptably low level.

Why is it that the negative hypergeometric smoothing method outperforms the other presmoothers? It is likely that it is in part because that smoother takes into account all of the information in a distribution's mean and standard deviation in arriving at the smoothed frequency for each point. Although the negative hypergeometric does require the assumption that all items are equally difficult, an assumption usually contradicted in practice, its success as a presmoother indicates that its use is robust against violation of this assumption. Furthermore, among the seven presmoothers investigated, only the negative hypergeometric is based on a mathematical model of testing.

The present study is limited in several respects, all of which may tend to reduce its generalizability to other applications.

First, only five tests were used: two operational and three simulated. Generalizations to other tests may be inadvisable if the tests do not statistically resemble those used for this study.

Second, the tests used, especially the simulated tests, may be more similar to each other than are most operationally equated tests. Generalization to less similar tests is of questionable appropriateness.

Among the presmoothing methods, the negative hypergeometric and, by extension, other smoothers of the same beta binomial family, deserve consideration for operational use. If any of the presmoothers studied here is to be adopted, then the negative hypergeometric would be the most appropriate. It has the effect of reducing RMSD by about ten percent, a benefit which could also be achieved by increasing sample size by about 20 percent.

References

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Belmont, CA: Wadsworth.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 508-600.

Holland, P. W., & Rubin, D. B. (1982). Test equating. New York: Academic Press.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika, 27, 59-72.

Lindsay, C. A., & Prichard, M. A. (1974). An analytical procedure for the equipercentile method of equating tests. Journal of Educational Measurement, 8, 203-207.

Lord, F. M. (1982). The standard error of equipercentile test equating. Journal of Educational Statistics, 7, 165-174.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1983). Comparison of IRT observed score and true score equatings. (RR-83-26-ONR), Princeton, NJ: Educational Testing Service.

Mosteller F., & Tukey, J. W. (1977). Data analysis and regression. Reading, MA: Addison-Wesley.

Ree, M. J. (1980). AVRAM: Adaptive vector and response automation method. Applied Psychological Measurement, 4, 277-278.

WEIGHTED LIKELIHOOD ESTIMATION OF ABILITY IN ITEM RESPONSE THEORY WITH TESTS OF FINITE LENGTH.

Thomas A. Warm, Ph.D.
U.S. Coast Guard Institute

A new method of statistical estimation, Weighted Likelihood Estimation (WLE), was discovered in Warm (1985a), and a new theorem of mathematical statistics was proved in Warm (1985b). The theorem states that WLE has zero first order bias, in contrast to Maximum Likelihood Estimation (MLE) and Bayesian Modal Estimation (BME) which are both biased. WLE was applied to ability (θ) estimation in Item Response Theory (IRT).

METHOD

Using Monte Carlo methods, WLE(θ) was compared to MLE(θ) and BME(θ) on 12 conventional tests with 10 to 60 items, a-parameters of 1 or 2, and normally distributed b-parameters. The three estimators were also compared on two tailored tests with optimal b-parameters. One tailored test had an infinite item bank and all $a = 2.00$. The other tailored test simulated a finite item bank with declining a-parameters. For all tests all $c = 0.20$.

RESULTS

Partial results are presented in the figures below. For complete results, see Warm (1985b).

In both conventional and tailored tests WLE(θ) was less biased than both MLE(θ) and BME(θ). In addition WLE(θ) had small variance over the entire range of the θ-scale, as well as small mean squared error even at non-central θ.

DISCUSSION

The relative unbiasedness of WLE(θ) makes this estimator particularly appropriate in applications of IRT for which the parameter invariance property is important.

Two new insights for MLE(θ) were discovered: 1) natural, rational bounds, and 2) a conditional analogy to the attenuation paradox in tailored tests with high a-parameters.

The heart of WLE(θ) is a weighting function, $w(\theta)$, which is multiplied times the likelihood function, and the product maximized. This weighting function, which removes the bias and uncontrolled variance of MLE(θ), is a function of θ and the item parameters, and is specific to each test. It was shown to be equal to the square root of test information for the one- and two-parameter models of IRT, and equal to a closely related function for the three-parameter model.

REFERENCES

Warm, T.A.(1985a) Weighted Likelihood Estimation of Ability in Item Response Theory. Technical Report CGI-85-01, U.S. Coast Guard Institute, P.O. Substation 18, Oklahoma City, OK 73169.

Warm, T.A.(1985b) Weighted Likelihood Estimation of Ability in Item Response Theory With Tests of Finite Length. Ph.D. Dissertation, University of Oklahoma, Norman, OK. (Also available as Technical Report CGI-85-08, U.S. Coast Guard Institute, P.O. Substation 18, Oklahoma City, OK 73169.)

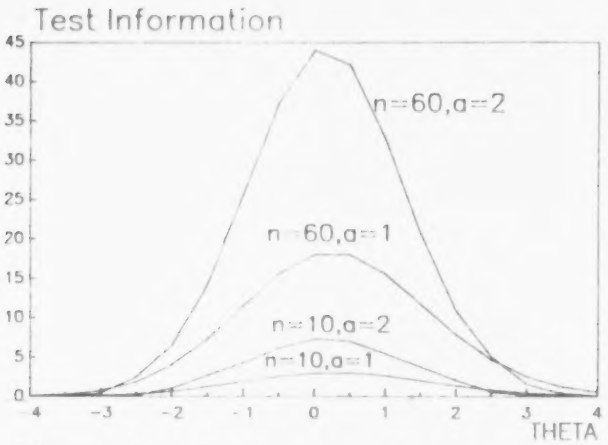


Figure 1. Test Information curves of the shortest (n=10) and longest (n=60) with all a=1 and all a=2.

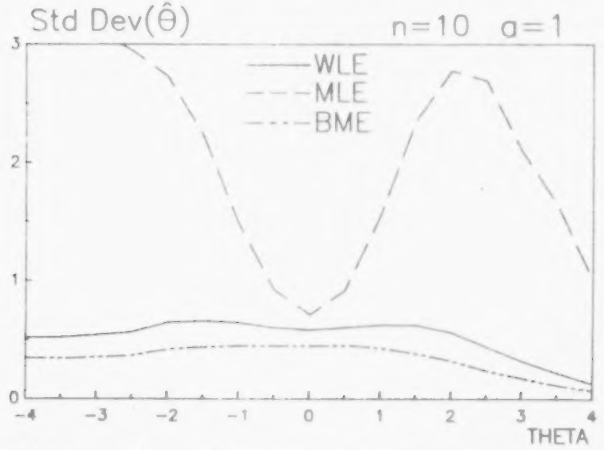


Figure 3. Standard Deviation of θ-estimates for conventional tests. SD(BME(θ)) was smallest, but SD(WLE(θ)) was also always small.

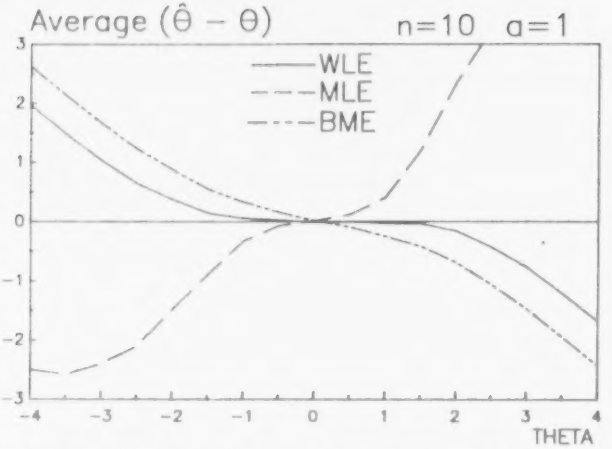


Figure 2. Bias of θ-estimates for conventional tests. WLE(θ) is less biased than MLE(θ) for all tests.

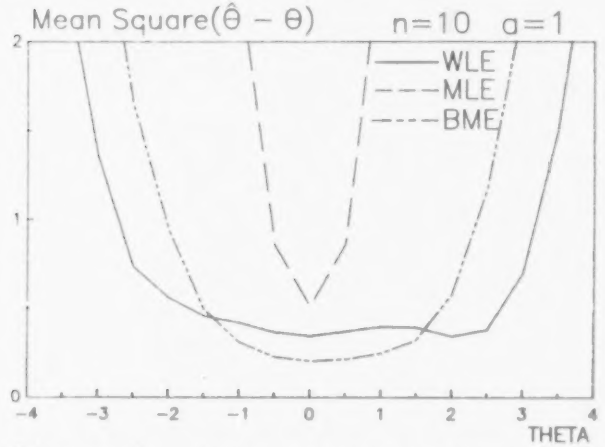
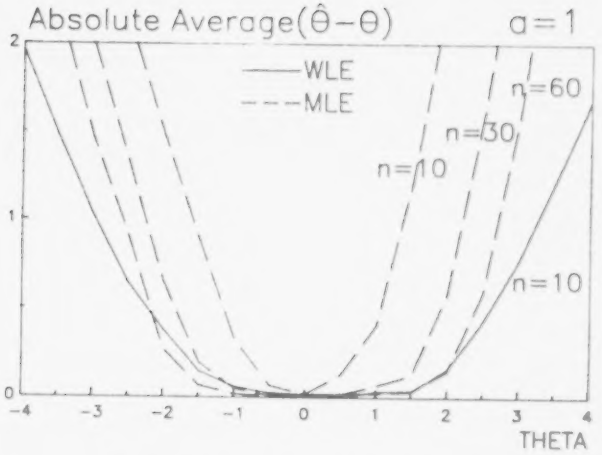
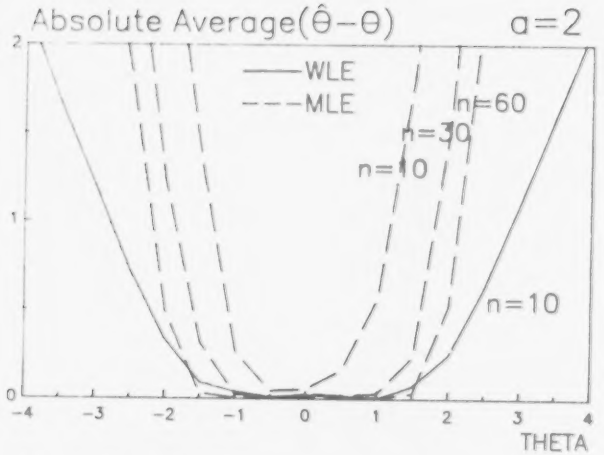
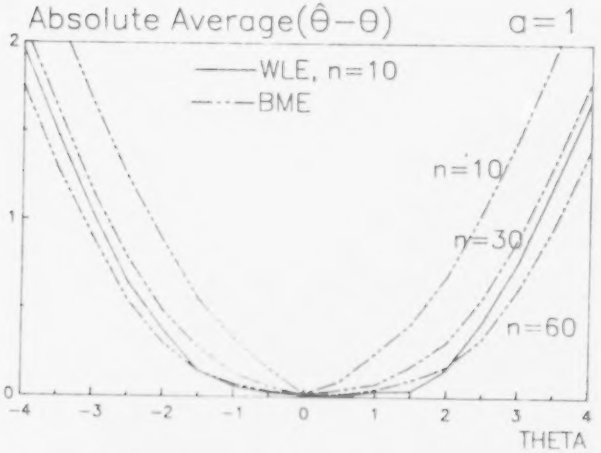


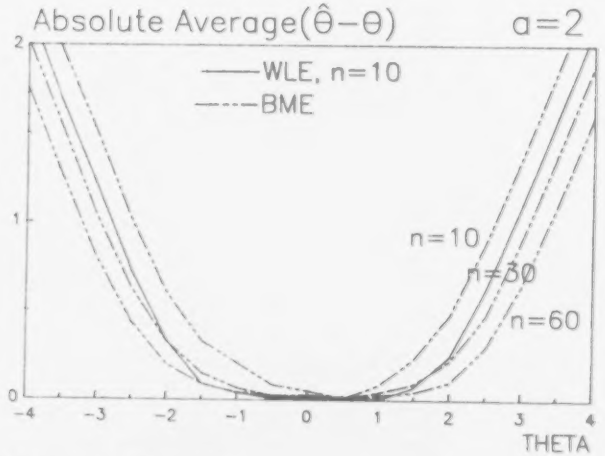
Figure 4. Mean squared error of θ-estimates for conventional tests. The range over which MSE(WLE(θ)) is small was always wider than for MLE(θ) or BME(θ).





Figures 5 and 6. Absolute bias of WLE(θ) and MLE(θ) on conventional tests. WLE(θ) with 10 items is less biased than MLE(θ) with 30 to 60 or more items.





Figures 7 and 8. Absolute bias of WLE(θ) and BME(θ) on conventional tests. WLE(θ) with 10 items is less biased than BME(θ) with 20 to 40 items.

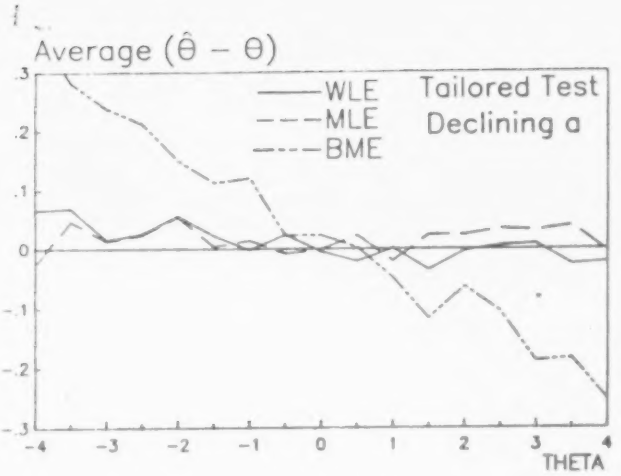
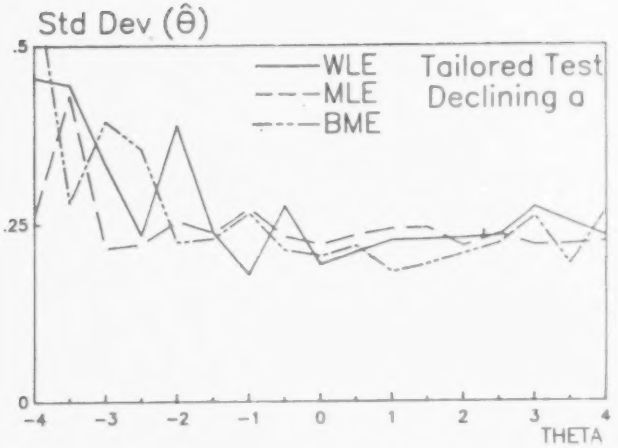
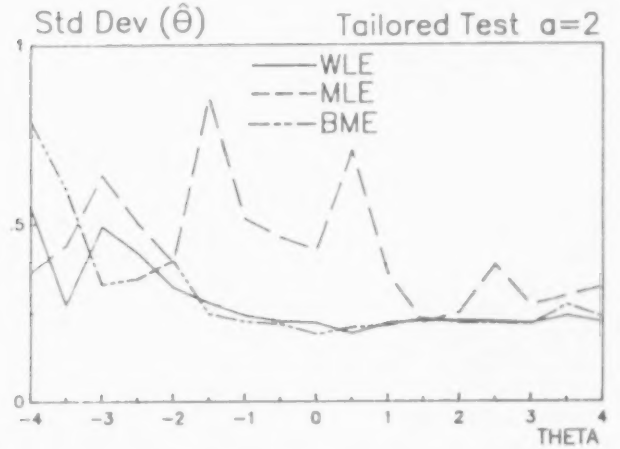


Figure 9. Bias of θ-estimates on tailored tests with simulated finite item bank. WLE(θ) is essentially unbiased. MLE(θ) is slightly biased, and BME(θ) is very biased.





Figures 10 and 11. Standard Deviation (SD) of θ-estimates on tailored tests with finite item bank(declining a-parameters), and infinite item bank(all a=2). With finite item bank SD of θ-estimates are about the same for all three estimators. Wtih infinite item bank MLE(θ) has high variance, exhibiting a conditional analogy of the attenuation paradox.

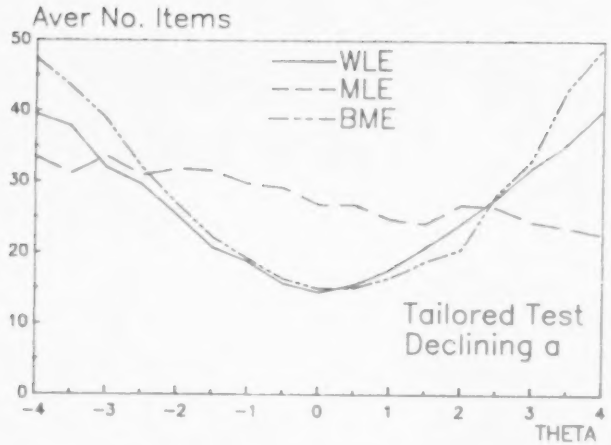


Figure 12. Average numbers of items administered for tailored tests. WLE(θ) used many fewer items than MLE(θ) at central values of θ.

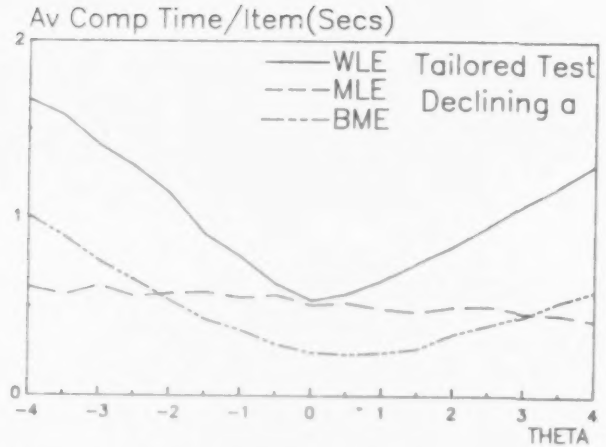


Figure 13. Average computation time (seconds) required between items administered by the tailored tests. WLE(θ) required two to three times as much computation time as either of the other two estimators.

LEADER BEHAVIOR AND THE PERFORMANCE OF FIRST TERM SOLDIERS

Leonard A. White, Ilene F. Gast and Michael G. Rumsey[1]
U.S. Army Research Institute for the Behavioral and Social Sciences

A large Army project is underway to validate new and current predictors of first term soldier performance. A major objective of this effort is to increase Army organizational effectiveness by improving the soldier job match. This will be accomplished by developing a set of selection and classification measures (predictors) and performance criteria and then empirically demonstrating relationships between the predictors and performance measures.

It is recognized that job performance is not only related to characteristics which are measurable and identifiable prior to enlistment, but is affected by experiences and developmental opportunities that occur throughout a soldier's life-cycle in the Army. The focus of the present research is on the performance-relevant consequences of a soldier's interaction with his or her superiors. Longitudinal research indicates that the quality of leader-subordinate work relationships are predictive of job success (Wakabayashi & Graen, 1984). Aspects of leader behavior such as providing rewards and recognition, disciplinary practices, and inspirational leadership have been related to subordinate effort and performance (e.g., Yukl, 1981).

However, past research on leadership and performance has generally omitted the influence of ability or the potential interactive effect between individual aptitudes and leadership on job proficiency and performance. Some investigations (e.g., Barnes, Potter, & Fiedler, 1983) have suggested that the prediction of job performance from general ability is moderated by leadership. Other researchers (Schmidt & Hunter, 1977) have argued that the relationship between general ability and performance is stable across time and situations for similar jobs.

To summarize, the model examined in this research assumes that job performance is influenced by a new incumbent's capabilities measured prior to enlistment and characteristics of the work environment. Within this framework the purpose of this research was twofold: (a) to examine relationships among dimensions of leader behavior and subordinate performance, and (b) to explore possible moderating effects of leadership on the correlation between general cognitive ability and job performance.

METHOD

Research participants were 696 first term soldiers in five military occupational specialties (MOS); 156 infantrymen (MOS 11B), 139 armor crewmen (MOS 19E), 125 radio teletype operators (MOS 31C), 141 light wheel vehicle mechanics (MOS 63B), and 135 medical care specialists (MOS 91A).

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

Of these soldiers, 88.5% were male and 11.5% were female; 28% were black, 3% were hispanic, 64% were white, and 5% other. Soldiers' report of work experience in their unit ranged from 2 months to 49 months (median=one year).

Instruments

The first step in this research was to develop measures of leader behavior and soldier performance on the job.

Supervisor behavior rating scales. Critical incidents workshops were conducted with 80 NCO in the five target MOS. These NCO generated a total of 474 examples of leader behaviors thought to influence soldier performance. Classification of the incidents by two of the authors and 31 NCO familiar with Army leadership requirements led to the identification of 9 categories of leader behavior (White, Gast, Sperling, & Rumsey, 1984). At least 5 and no more than 8 items were written to represent important leader behaviors in each category (e.g., Your supervisors are hard to find when you need them). These procedures resulted in a 60-item question naire. Responses to each item were made on a 5-point scale from very seldom or never (1) to very often or always (5).

Job performance rating scales. To develop these scales, critical incident workshops were conducted in which NCO provided examples of effective (as well as ineffective) soldier performance. The number of NCO and examples provided were as follows: MOS 11B, 51 NCO's, and 906 incidents; MOS 19E, 43 NCO's and 798 examples; MOS 31C, 45 NCO's and 830 incidents; MOS 63B, 49 NCO's and 882 incidents and; MOS 91A, 42 NCO's and 783 incidents. A variant of the behaviorally anchored rating procedure (Smith & Kendall, 1963) was used to develop behavior-based rating scales for each job. The resulting rating form for each job consisted of seven to ten 7-point behavior summary scales.

Army-wide performance rating scales. To prepare these scales, 77 NCO's and junior officers working in a wide variety of Army jobs generated 1,215 behavioral examples. The examples represent those aspects of soldier effectiveness that contribute, broadly speaking, to organizational effectiveness, such as following orders and regulations. The target criterion space for these scales went beyond job performance to include aspects of socialization and commitment to the organization. Eleven 7-point behaviorally anchored rating scales were developed for each job.

Hands-on, task proficiency tests. For each of the jobs, 5-8 critical tasks were identified to represent the MOS-specific task domain. Multistep task proficiency tests were prepared for each task. Each step of a task was scored pass or fail. A score for each task was computed by calculating the proportion of steps passed and the task scores were averaged to yield an overall hands-on test score.

Job knowledge tests. Through job analysis, important knowledge areas were identified for each of the five jobs. With the help of subject matter experts, items were written to tap these knowledges. For each soldier, the percentage of correct items was the overall job knowledge test score.

Correlations of hands-on and job knowledge test scores, job performance ratings, and the Army-wide effectiveness rating with the leader behavior scales are presented in Table 2. Results are shown separately for each of the five jobs. A mean correlation ($\bar{r}$) across the five jobs was computed by weighting each correlation by its associated sample size (Hunter, Schmidt, & Jackson, 1982). The highest correlations were

Table 2

Correlations between Leadership Scales and Criterion Measures by Army Job

	Leadership Scale								
Job	1	2	3	4	5	6	7	8	Total Scale
Hands-on Task Proficiency									
11B	.05	.03	.19*	.24*	.17*	.25*	.10	.11	.18*
19E	.14	.11	-.01	.14	.24	.11	.18	.23	.22
31C	.02	.00	.15	.18	.02	.08	.03	.15	.09
63B	.12	-.05	.10	.06	.00	.05	.08	.12	.09
91A	-.05	-.14	-.12	.02	-.04	-.08	-.09	-.08	-.10
$\bar{r}$	.05	-.02	.08	.13*	.07	.09	.05	.10	.09
Job Knowledge									
11B	-.01	-.17*	.02	.13	.03	.15*	.09	.12	.03
19E	-.03	-.03	.05	.03	.06	-.13	-.02	-.03	-.02
31C	.17	.11	.12	.30*	.26*	.23*	.12	.17	.22
63B	-.05	-.04	.05	.05	.20*	-.05	.01	-.06	-.01
91A	-.12	-.10	-.01	-.01	-.11	-.18	-.22	-.23	-.13
$\bar{r}$	-.01	-.06	.04	.09*	.08	.00	.00	-.01	.01
Job Performance Rating									
11B	.12	.01	.05	.23*	.08	.21*	.10	.03	.13
19E	.11	.04	.09	.16	.08	.05	.11	.21	.13
31C	.21*	.01	-.04	.12	.20*	.17	.02	.07	.14
63B	.17	.08	.23*	.08	.20*	.07	.07	.08	.18*
91A	.08	.07	.05	.11	.00	.08	-.12	.01	.03
$\bar{r}$	.13*	.04	.08	.14*	.11*	.11*	.04	.08	.12*
Army-wide Effectiveness Rating									
11B	.17*	.06	.04	.23*	.12	.20*	.11	.07	.17*
19E	.12	.02	.07	.14	.22*	.10	.14	.15	.15
31C	.41*	.19*	.12	.34*	.32*	.32*	.18*	.24*	.37*
63B	.20*	.13	.30*	.11	.19*	.06	.04	.08	.21*
91A	.17*	.05	.14	.20*	.07	.09	-.09	.07	.12
$\bar{r}$	.21*	.09*	.13*	.20*	.18*	.15*	.07	.12*	.20*

Note. Leadership scales: 1 (Support); 2 (Informing); 3 (Fairness); 4 (Participation); 5 (Performance Contingencies); 6 (Role Clarification); 7 (Results Orientation); 8 (Training & Development); 9 (Total).

*$p<.05$

General cognitive ability. The Armed Services Vocational Aptitude Battery (ASVAB) was administered to all participating soldiers prior to entering military service. The ASVAB, which consists of ten subtests, is used for selection and occupational classification. A composite measure of four ASVAB subtests, known as the Armed Forces Qualification Test (AFQT), was used as the measure of general cognitive ability.

Procedure

Raters were trained to use the behavior-based rating scales. After training, supervisors in groups of 3-15 evaluated their subordinates on the Army-wide and job performance rating scales. The mean number of supervisor raters/ratee ranged from 1.66-1.83 for the five MOS. Ratings were averaged across supervisor raters to form an overall job performance rating and an Army-wide effectiveness rating for each ratee.

The first term soldier (ratees) completed the supervisor behavior rating scales, and were also administered tests of job knowledge and hands-on, task proficiencies.

RESULTS

Principal components factor analysis was used to examine the dimensionality of the supervisor behavior rating scales. Varimax and promax solutions were computed and the interpretation restricted to factors appearing in both solutions. Comparison of the rotated structures yielded eight factors with eigenvalues greater than one. Items loading above .4 on one and only one factor were interpreted as measuring the factor. Items with weak loadings on all factors or similar loadings on two or more factors were not used to measure any factor. Factor score estimates were computed by unit weighting and summing individual's responses to the set of items representing each factor. Table 1 presents the intercorrelations among the estimated factor scores.

Table 1

Intercorrelations Among Leadership Scales and Summary Statistics.

										No. of Items	Scale Mean	Std. Dev.
1. Support/Inspiration	.89	.64	.48	.64	.53	.72	.58	.68	.90	9	27.2	7.5
2. Informing		.78	.54	.49	.53	.54	.48	.50	.79	6	19.0	4.8
3. Fairness			.74	.47	.46	.40	.31	.36	.67	5	16.5	4.3
4. Participation				.70	.44	.60	.46	.56	.76	4	13.4	3.4
5. Performance Contingencies					.55	.47	.40	.39	.67	3	9.9	2.5
6. Role Clarification						.78	.55	.63	.80	4	12.9	3.1
7. Results Orientation							.56	.59	.66	3	9.4	2.2
8. Training and Development								.72	.77	5	14.7	3.9
9. Total									.94	39	123.1	24.7

Note. Internal consistency reliabilities are presented on the diagonal.

$^{a}N = 696$

obtained between perceptions of leader behavior and the Army-wide effectiveness ratings. Within the set of Army-wide performance dimensions, strongest relationships were obtained between supportive and participative leadership and ratings of subordinate adherence to regulations and willingness to provide extra effort when needed. Statistically significant but low correlations between leader behaviors and job proficiency were evident in the two combat MOS.

Hierarchial regression analysis was used to estimate the relationships of cognitive ability (i.e. AFQT score), leadership climate, and their interaction to job proficiency and performance. The AFQT score was entered first in the regression to assess the contribution of mental ability at the time of enlistment to later job performance. Then, leadership and the ability X leadership interaction were entered to assess post-enlistment leader influences on performance and the utilization of ability on the job. In the regression model, leadership was represented by the sum of scores on the 8 leadership scales. The criterion variables were job knowledge, hands on task proficiency, and supervisor ratings of job performance and Army-wide effectiveness.

Of interest here, results of the regression analyses revealed no statistically significant increase in $\underline{R}^2$ due to inclusion of the ability X leadership interaction in the model. In each of the five jobs, the highest multiple correlations were obtained for prediction of job knowledge, with $\underline{R}=.30$, to .60, all $\underline{p}<.05$. This effect was primarly attributable to the influence of general ability on job knowledge. Leadership and cognitive ability had significant independent effects on task proficiency in the infantryman and armor crewman jobs with, respectively, $\underline{R}=.28$, $p<.05$, and $\underline{R}=.37$, $p<.05$. However, in MOS 91A and MOS 63B $\underline{R}^2$s for the prediction of task proficiency from the independent variables failed to reach significance. With respect to supervisory ratings of job performance, ability and leadership and their interaction accounted for less than 5% of the variance in this criterion. Leadership showed several significant correlations with Army-wide effectiveness ratings at the zero-order level, however the $\underline{R}^2$ for this criterion achieved significance only in the radio-teletype operator job, with $\underline{R}=.37$, $\underline{p}<.05$. Correlations between cognitive ability and the Army-wide effectiveness rating ranged from $\underline{r}=-.28$ to .03.

DISCUSSION

The present research explored relationships between leadership, cognitive ability, and the performance of first term enlisted soldiers. Results for the five Army jobs examined here support the conclusion that general ability and leadership behavior have independent effects on performance. However, each appears to contribute to effective soldiering in different ways. Leadership, as perceived by the subordinate, had the strongest effect on the motivation-related, dependability facets of performance measured by the behaviorally based rating scales. General cognitive ability contributed to performance by enabling enlistees to learn the facts and procedures required to perform their jobs.

No evidence was obtained indicating that relationships between general ability and job proficiency and performance are moderated by leadership influences. This finding supports conclusions by Schmidt and

Hunter (1977) that the validities of cognitive tests are similar across situations for the same job. Correlations between general cognitive ability and each criterion measure did vary somewhat across jobs, but almost all of the variation was attributable to sampling error.

The relationships between leadership and performance reported here should not be interpreted as indicating that leadership behavior "causes" performance. Leadership effects on performance may be understood in terms of exchange theory (Graen, 1976) which views the interaction between leader and subordinate as a reciprocal influence process that develops over time. Subordinates who are perceived as willing to work hard and support the mission will be evaluated more favorably by their superiors. In return for their support, these soldiers are likely to receive more individualized attention, information, and other resources from their supervisors; which, in turn, serves to reinforce and sustain subordinate effort.

The results reported here are largely exploratory. Future data collection and analysis will provide an opportunity to confirm the leadership factors and to examine potential moderating effects of leadership behavior on a broad range of soldier aptitudes and characteristics.

REFERENCES

Barnes, V., Potter, E. H., & Fiedler, F. E. (1983). Effect of interpersonal stress on prediction of academic performance. Journal of Applied Psychology, 68, 686-697.

Graen, G. (1976). Role-making processes within complex organizations. In M. Dunnette (ed.) Handbook of Industrial Organizational Psychology. Chicago: Rand McNally.

Hunter, J.E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating results across studies. Beverly Hills: Sage Publications.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

[illegible] M., [illegible] (1984). The Japanese career progress study: A 7-year follow-up. Journal of Applied Psychology, 69, 603-614.

White, L. A., Gast, I. F., Sperling, H. M., & Rumsey, M. G. (1984, Oct). Influence of soldiers' experiences with supervisors on performance during the first tour. Paper presented at meeting of the Military Testing Association, Munich, Germany.

Yukl, G. A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

MEASURING MILITARY HEROISM

Jeffrey W. Anderson
U.S. Army Research Institute

To a large extent the drama of human history has been woven from the biographies of people who by the nature of their successes beyond the pattern of everyday life assumed a mythical or legendary stature and became national heroes (Downton, 1973). During periods of revolutions and wars the fate of entire peoples has seemed to hang upon the decisions of a few powerful people. Yet, Arthur M. Schlesinger, Charles de Gaulle, and others have remarked that the age of heroes is past (Jennings, 1960). Yet it is also possible that the character of our heroes, or what we recognize as heroism, has changed. We no longer regard our heroes as rare specimens of humanity. They are only somewhat different from ourselves generally due to the situation or circumstances surrounding the act of heroism. While gallant acts have given birth to legends of heroes, scientific research in the area of heroism has been very scarce. The present study is a preliminary attempt to examine four research questions. First, what are the ingredients of heroism? Second, do people still recognize and label certain behaviors as heroic? Third, have these heroic characteristics changed over time? And fourth, can we measure heroism before a crisis?

Heroism involves great bravery, daring coufage, valor, gallantry, and intrepidity (Barnhart, 1967). It includes elements of extreme self-sacrificing courage, fulfilling a high purpose, and attaining a noble end (Webster, 1974). But these qualities do not readily lend themselves to measurement and research. In the literature on heroism there are three major foci. The first emphasizes the personality of the hero. It contends that the hero would have been socially recognized as a hero without regard for the circumstances surrounding his behavior. The second focus follows the arguments of Hegel and Spencer. Man is considered a mere product of social forces coincidentally combined in time and space to converge upon an individual that society labels as a hero. Finally, early social reformers and revolutionaries, emphasizing Darwinian concepts, contended that heroes were thrown up by some chance of the natural selection process. The social environment was a selection instrument to provide opportunities for these men to display their hereditary talents in heroic acts. The situation limited the hero, but did not dominate him.

In general, the literature of heroism tells us little about the type of person who is predisposed, with situation permitting, to become a military hero. In other words, while we all seem to know about heroism there is no operational definition of the concept. To attempt to provide an operational definition, we systematically analyzed the citations for heroism in which the recipient was given the highest military award, the Medal of Honor. These citations should also represent our society's general definition of heroic action. The Medal of Honor is presented to a service member who "distinguished himself conspicuously by gallantry and intrepidity at the risk of his life above and beyond the call of duty." As the highest military award for bravery, each branch of the armed forces has established a set of prescrip-

---

The views expressed in this paper are those of the author and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

tive regulations that leave no margin of error or doubt. The deed of the recipient must be proved by incontestable evidence from at least two eyewitnesses, and it must be so outstanding that it clearly distinguishes his gallantry beyond the call of duty from lesser forms of bravery. As such, the citations from Medal of Honor winners are essentially narratives of critical incidents of heroism and well-suited to the development of an operational definition of military heroism (Flanagan, 1954).

Method

At the time of this study the Medal of Honor had been awarded 3369 times, spanning the years 1863 to 1978. From the citations given with these awards, 337 were randomly selected for analysis.

A group of three judges, two male and one female, working independently, analyzed these citations to determine the critical behaviors exhibited by the hero that were recognized by his fellow s 'iers and rewarded by the military. The eight dimensions of behavior that were unanimously listed by all three judges were accepted as truly describing military heroism.

Another group of five judges, two female and three male, were then asked to rate each of these dimensions concerning how well it described military heroism as exemplified by a group of 254 randomly selected different citations. These additional citations were chosen in an effort to broaden the sample and thereby improve the generalizability of the dimensions and findings. In order to achieve these objectives ten percent of the remaining 3032 citations or 303 citations were randomly selected. Since many of these citations were verbatim duplicates of others, redundant citations were elkminated, giving 254 citations for further study. These judges had extensive military experience and represented the Army, Navy, and the Air Force. Each judge was asked to rate on a 1 to 5 scale the extent to which the derived eight dimensions described the behavior of the award recipient in each citation (from not at all descriptive to totally descriptive).

Since the judges used the rating scales differently, each rating score was converted to a standard score using the overall mean and variance of the judge giving the rating. The interrater reliabilities, using Winer's technique and corrected for five judges, were calculated. Based on these reliabilities, the ratings from all five of the judges were combined and a mean rating was calculated for each performance dimension. The more accurately a dimension described the concept of heroism as defined by a specific citation, the higher the judge's rating on that dimension. The dimensions were ranked in order of their overall importance in describing military heroism.

Subsequently, ratings were categorized according to the conflict from which the Medal of Honor citation was drawn. A comparison between conflicts of relative importance of each dimension to the concept of Military Heroism was used to test the implied hypothesis that our societal definition of heroism had changed over time.

Results

The first group of three judges were asked to independently derive dimensions of military heroism based on a randomly selected sample of 337 citations from the Medal of Honor. Dimensions that were unanimously selected by the judges were used to describe each citation. All citations could be described using one or more of the dimensions presented in Table 1.

Table 1. Dimensions of Heroism in order of importance with Interrater Reliabilities

Dimension		r
The hero is thoroughly devoted to accomplish his duty.	(Devotion to Duty)	.47
The hero sets a personal example of behavior for others.	(Personal Example)	.45
The hero risks his own life or places himself in danger.	(Accepting Danger)	.76
The hero rescues or saves another person.	(Saving Life)	.92
The hero overcomes his own injuries or illness.	(Overcoming Injury)	.93
The hero succeeds when the odds are overwhelmingly against him.	(Defeating Great Odds)	.71
The hero takes command or gives leadership when it is lacking.	(Taking Command)	.73
The hero seizes upon an opportunity.	(Seizing an Opportunity)	.57

Essentially, the military hero had to set an example of behavior before and after formal recognition by the organization. He persisted in the accomplishment of his duty and willingly accepted personal danger, subordinating his own life to the values of his cause. Given the opportunity to command (leadership) he did so, and when opportunity presented itself he seized upon it.

These dimensions did not, however, demonstrate any relative importance in describing heroism. A second set of judges, therefore, was asked to rate on a scale of 1 to 5 how important each dimension was in describing the heroism expressed by each of 254 different, randomly selected citations. After converting the ratings given by each judge to standard scores, the interrater reliabilities for each dimension, using Winer's technique and corrected for five judges, were calculated. These reliabilities are also presented in Table 1. While the first two dimensions show lower reliability than other dimensions, all are considered acceptable for untrained raters, especially in light of the unanimous agreement given to these two dimensions by the original three judges.

Based on these reliabilities, the judges' evaluations were combined. The results of this combination yielded the order of importance for the dimensions as shown in Table 1. In all cases, the five judges agreed that the dimensions previously constructed to describe heroism were adequate descriptors of heroism.

To test the implied hypothesis that military heroism has changed over time, each of the judges' ratings were grouped according to the time period of the citation from which the rating was derived and average ratings for each dimension were calculated for seven major conflict periods from the Civil War through the Vietnam conflict. A profile analysis of these average ratings by conflict period is shown in Figure 1.

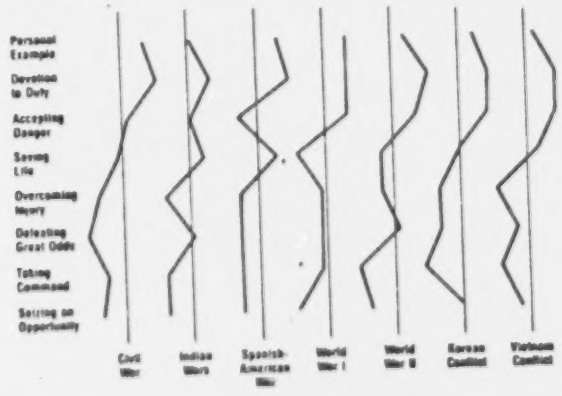


Figure 1. Profile Analysis of Ranks by Conflict

If the center line above each conflict is viewed as the mean, deviations to the right indicate that the dimension described the citations for heroism during that conflict, while deviations to the left indicate that the dimension was not descriptive of the military heroism citations during that conflict. The first three dimensions listed were most often chosen by the judges as describing military heroism in every conflict period. This analysis suggests that the same dimensions used by military leaders to describe heroism during the Civil War apply to heroism recognized in Vietnam and implies that military heroism as a construct is relatively stable over time.

To support further the argument that the first three dimensions are interrelated and describe heroism, the intercorrelation matrix was computed. The correlations between the first three dimensions are three of the four highest in the computed matrix indicating that indeed these dimensions are highly interrelated. Further, a factor analysis of the eight original dimensions with a Varimax rotation is presented in Table 2. The eight descriptors may be reduced to two factors with eigenvalues above 1.0. Again the first three dimensions load highly on the first factor (loadings of .60 and above). The first factor containing the first three dimensions appears best for describing heroism. The second factor seems to be a measure of the individual's ability to recover a deteriorating situation. While this is certainly commendable behavior, the judges seem to indicate by both their ratings and their frequency of selection of dimensions that this is not worthy of the term heroism. Further, this factor has an eigenvalue only slightly above 1.0 and seems, therefore, to be less reliable than the first factor.

Table 2. Factor Analysis

Factor 1.	
Item	Loading
Personal Example	.60
Devotion to Duty	.79
Accepting Danger	.73
Overcoming Injury	.44
Eigenvalue	3.58

Factor 2.	
Item	Loading
Saving Life	.31
Defeating Great Odds	.50
Taking Command	.48
Seizing an Opportunity	.95
Eigenvalue	1.09

Discussion

From the analysis presented, it appears that there are certain common elements in the dimensions used by the original three judges to describe heroism. They substantiate that the first three dimensions describe military heroism and then form the operational definition. The hero sets an example of personal behavior for others in accomplishing his purpose or duty, which is so valuable that he willingly endangers his own life to complete the task.

A combination of these three dimensions of military heroism could also describe the Warrior Spirit sought by the Army. The Warrior Spirit encompasses all of the physical, mental and moral qualities essential to successful soldiers in combat. Based on an analysis of military history, in a follow-on effort we described the Warrior Spirit as:

1) A selfless devotion to accomplishing a duty or perceived noble cause.
2) Leadership by personal example--especially applying high but achievable standards to himself and his unit.
3) A reasoned acceptance of risk (especially risk of his own life)--calm, confident and self-controlled in the face of mortal danger.
4) Decisiveness despite unreliable, incomplete and often inaccurate information (ability to separate the important from the trivial).
5) Being effective at communicating instructions so that every member of the unit knows and understands the leader's wishes.
6) Creating a team or cohesive unit that all work as one to achieve the noble cause or purpose and training that unit for combat.

The Warrior Spirit, then, appears to be a combination of native characteristics and training. The Fighter Studies (HumRRO, 1957-1958), examined effective combat soldiers during the Korean conflict. They found that a fighter (warrior) tended to (1) be more intelligent, (2) be a "doer", (3) have greater emotional stability, (4) have better health and vitality, and (5) have a greater fund of military knowledge. In 1979 and 1980 studies, Anderson found that successful combat leaders were more intelligent, task-oriented, had higher morale, and had more direct, job-related experience than their less successful peers.

An historical analysis by the Department of History, USMA (1984) found that successful combat leaders had: (1) terrain sense, (2) single-minded tenacity - moral courage (3) ferocious audacity - willing acceptance of reasoned risk, (4) physical confidence, and (5) practical, practiced judgement - common sense.

Based on these studies the hero may be selected in peacetime based on his intelligence, moral courage, character, mental and emotional health, physical well-being (medical and athletic), decision-making ability, common sense, and self-confidence.

Conclusions

This paper has presented a simple, empirical investigation to determine the operational definition of military heroism. It has corroborated this definition with findings from other, related research. Though current psychological literature has little information concerning the psychological profile of a hero, there are certainly indications from the analysis of a different body of literature that military heroism is readily recognizable for others and that the hero has certain measurable characteristics that distinguish him from common humanity. We have shown the underlying factors of heroism and in conjunction with related research propose that certain factors may be measu e an used to predict military heroism. Since this study is the first of its kind, there are admitted imperfections, but it demonstrates conclusively that heroism has a psychological meaning for the average individual which may be scaled and reliably measured.

REFERENCES

Anderson, J. W. (1980). The prediction of combat effective leadership (Doctoral dissertation, University of Washington, 1980). Dissertation Abstracts International, 41, 1968b.

Barnhart, C. L. (Ed.) (1967). Comprehensive desk dictionary. Garden City, NY: Scott, Foresman, and Company.

Committee on Veterans Affairs. (1979). Medal of Honor recipients, 1863-1978. Washington, D.C.: Author.

Department of History. (1984). Leadership in combat: An historical appraisal. Unpublished manuscript, United States Military Academy, West Point, New York.

Downton, J. V., Jr. (1973). Rebel leadership. New York: The Free Press.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Human Resources Research Office. (1957). Fighter I: An analysis of combat fighters and non-fighters (HumRRO Technical Report No. 44). Washington, DC: Human Resources Research Office. (DTIC No. AD 158 148).

Jennings, E. E. (1960). An anatomy of leadership. New York: Harper and Brothers.

Webster's new collegiate dictionary. (1974). Springfield, MA: G. & C. Merriam Company.

ANALYSIS OF FUNCTIONAL TRAINING PROGRAMS AT THE NAVAL SUPPLY CENTERS

Neal R. CROWLEY
U. S. Naval Audit Service

In fiscal 1980-1981 $223,649,000 was misplaced, lost, or stolen from the Naval supply centers and management indicators tracking worker performance declined. Congressional concern resulted in several General Accounting Office (GAO) audits and numerous recommendations for improvement. In fiscal 1982 performance indicators improved and $48,974,000 in misplaced material was located.

Responding to strong Congressional and GAO criticism for poor inventory accuracy the Naval Supply Systems Command (NAVSUP) instituted several programs to improve the performance of warehouse workers. As part of the overall effort, NAVSUP started a functional training effort called Competency Based Certification (CBC). The CBC program increased training staffs, improved training facilities, provided money for more training, required the supply centers to develop job related training materials for their warehousing functions and to fully train all employees. CBC required supervisors to "certify" that employees could perform their jobs.

Scope of the CBC Program

Approximately 1,500 warehouse workers at seven supply centers received 79,776 hours of CBC training from 1 August 1982 until 31 December 1984. NAVSUP spent almost $3,500,000 during this time for contractor developed CBC training materials and $500,000 for classroom renovation, printing, etc. NAVSUP hired 45 new training employees to implement the program and issued a new $12,000,000 training contract because of management's belief in the success of the program.

CBC training is very expensive. Contractor costs alone exceed $5,000,000 so far for a per person cost of over $3,333. 79,776 hours of CBC training equals slightly over 53 hours per employee. NAVSUP paid an additional $970,000 for employee salaries while they were being trained and almost as much for the salaries of program administrators and instructors. Spread over the 27 months of the program studied, 53 hours of CBC training gave each employee an average of two hours per month.

Currently, more hours go into administering the program and developing training materials than go into employee training.

There are two training programs. One is based on contractor and supply center training materials. Another is physical distribution training. This consists primarily of greatly intensified training along traditional lines. Employees go to existing training courses ranging in size from one hour briefings to six month professional training courses. The supply centers conducted 226,429 hours of physical distribution training at a salary cost of $2,750,000. The cost of travel, per diem, tuition, materials, etc., for physical

distribution training was approximately $500,000. Administrative costs were small.

This evaluation concentrates on the time spent on training and its effect on worker quality, timeliness, and productivity. Has the CBC program worked? To find out I analyzed the program at all seven supply centers. The basic question is whether the amount of training is related to supply center performance as measured by management indicators. If training changed performance, then the change must show up in performance and productivity indicators.

Although assessing the effectiveness of training is the main objective, I also evaluated the relative efficiency of the two training programs.

The primary hypothesis is that there is a strong, or at least moderate, statistical relationship between training and performance. A secondary hypothesis suggests that the relationships between quality, timeliness and productivity can be explained or depicted in a logical and statistically significant way.

Variables Used

I conducted a time series analysis looking for inter-relationships between nine variables: quality, timeliness, both quality and timeliness (Q & T), work unit productivity, receipt and issue productivity, overtime, each type of CBC training, and a combination of all training. The quality variable measures how accurately warehouse workers perform their jobs. The timeliness variable measures how quickly workers process materials. The Q & T variable measures both accuracy and speed of worker performance. Two variables measure worker productivity and one measures overtime.

I used four types of analysis: first looking for a statistically significant change in performance; second evaluating the change by comparing the seven supply centers; and linear and multiple regression analysis for each supply center.

How I Computed the Variables

Three variables measure the accuracy and speed of warehouse worker performance. The means of five indicators make up the quality variable for worker performance. The means of six indicators make up the timeliness variable for worker performance. The composite mean for all eleven indicators makes up the Q & T variable. I took work unit and manhour statistics from official documents.

Threats to Internal Validity

Several factors threaten the internal validity of the analysis. Norfolk changed its receipt and issue testing procedures to stop counting previously recorded errors. NAVSUP program managers caught two supply centers fudging results.

Cheating may be widespread. Some data is missing. But, overall, the data used appears adequate.

Changes in procedures, policies, and equipment take place constantly at each supply center. The use of seven supply centers as comparison groups helped in isolating the probable cause of changes.

The turnover rate varies at each supply center and this undoubtedly influences performance. Unfortunately NAVSUP does not keep turnover rates by department or type of work.

Small Sample T Test

I used a standard small sample t test to evaluate the change in performance before and after training for the six non-training variables. The null hypothesis is that the difference between the means before and after training is due to chance. The alternative hypothesis is that a significant change took place. On the basis of a one-tailed test at the .01 level of significance, I would reject the null hypothesis if a t were greater than t.99, which for 12+9-2 = 19 degrees of freedom is 2.539. On the basis of a one-tailed test at a .05 level of significance, I would reject the null hypothesis if t were greater than t.95, which for 19 degrees of freedom is 1.729.

Overall, there is a significant difference between the means before and after training. I therefore rejected the null hypothesis. I evaluated the total change in performance for each supply center and conducted linear and multiple regression analysis to see if the change is related to training.

Cumulative Change in Performance

I computed the cumulative change in performance since the CBC program began for each supply center and then used linear regression analysis with the change in performance as the dependent variable and training as the independent variable. A minimum of F=6.61 is required for a confidence interval of 95%. F Test results have five degrees of freedom.

The correlation coefficients between CBC training and changes in performance are very low. CBC training did not come close to passing the F test for any performance indicator. The hypothesis requires that the more training, the greater the improvement in performance. The null hypothesis maintains that training and changes in performance are unrelated. This analysis fails to reject the null hypothesis for CBC training.

Physical distribution training shows a much stronger correlation with changes in performance than CBC training does. However, the only relationship with training that is statistically significant at the 95% probability level is receipt and issue productivity. Overall, the combination of CBC and physical distribution training shows a stronger correlation than either of the two alone, especially for timeliness and Q & T.

There is a statistically significant inverse relationship between overtime and quality. Even the corrected coeficient of determination is an impressive .79. The T test result is -4.8 giving it a probability of less than .005 that the result is due to chance.

Overtime appears to be more directly correlated with performance than training. As the amount of overtime declines the quality of performance improves. Inversely, as the amount of overtime increases quality declines. The relationship between overtime and timeliness is mixed.

Linear Regression for Each Supply Center

I conducted linear regression analyses and obtained the following results.

Jacksonville's, Norfolk's and Oakland's F test scores for CBC's correlation with receipt and issue productivity exceed the 5.59 required for a probability of 95%. No other CBC correlation is significant. There may be a causal relationship between changes in productivity and CBC training, that is, the more training, the better the performance for Jacksonville, Norfolk, and Oakland but not for Charleston, Pearl Harbor, Puget Sound, or San Diego.

Puget Sound shows a significant correlation between physical distribution training and quality. Jacksonville shows a significant relationship between physical distribution training and receipt and issue productivity.

Oakland shows a strong correlation with work unit productivity and Puget Sound shows a strong correlation with quality. Overall there was a slight improvement with CBC training removed except at Norfolk which shows a slight decline.

Overtime shows the strongest correlations with performance. Norfolk's overtime is related to timeliness, receipt and issue productivity, and work unit productivity with a probability of error of less than .05%. By far the strongest relationship (a probability of 99%) is at San Diego for timeliness, and Q & T.

General Trends

General trends are mixed. Overtime shows an overall negative correlation with performance but a mixed negative and positive correlation with productivity. Overtime work hours, result in higher production but overtime also increases the overall number of hours required to do a job and results in lower productivity when the relative portion produced during overtime falls below the relative portion produced during regular work time.

Both productivity measures show a mixed negative and positive correlation with all performance indicators and a positive correlation with timeliness indicators. Work unit productivity also shows a positive relationship in this area. CBC and physical distribution training show a weak overall

inverse relationship with productivity, quality, and timeliness.

Multiple Regression

I conducted 57 multiple regression analyses for each supply center. Mixed results make interpretation difficult but do suggest some important relationships.

The dependent variables were quality, timeliness, Q & T, and the two measures of productivity. The strongest correlation overall exists between overtime and training as independent variables and performance measures as the dependent variables. Productivity (or the amount of work done) when used as an independent variable, is negatively correlated with quality and positively correlated with timeliness, and Q & T. Training is positively correlated with performance, except at Oakland where the relationship is negative for all training and Puget Sound where the relationship is nagative for CBC training.

Overtime shows a strong inverse relationship to performance indicators. In the case of quality this makes sense. It suggests the more hours people work the more mistakes they make. It is not clear why timeliness and productivity have inverse relationships with overtime. Next to overtime, all training shows the strongest correlation with various dependent variables. Overall, it is more effective than either CBC or physical distribution training alone. CBC'c inverse relationships suggest a problem with program administration. In some cases CBC training was poorly planned and executed and adversely affected performance. The time devoted to CBC training may have detracted from performance and productivity.

Training's strongest influence is on productivity. Four CBC, three physical distribution, and six combined training t tests exceed the 90% level of significance when used to test the relationship with productivity. One additional CBC, one physical distribution, and one combined test were negative at the same level.

Both types of training correlate with timeliness. There is a strong relationship between productivity and timeliness. Overtime is also highly correlated with timeliness. Puget Sound appears to have a strong physical distribution training program. This could be responsible for improvements in performance. Other than Puget Sound, and discounting Oakland's inverse relationship, there is not a significant correlation between training and quality or timeliness of performance for any supply center.

At Charleston productivity is inversely related to quality and overtime is inversely related to work unit productivity. The only significant result for Jacksonville is between training and productivity. Norfolk shows an inverse relationship between overtime and productivity and a positive relationship between overtime and timeliness. CBC training is also positively related to productivity and Q & T.

Oakland's the relationship between training and performance is generally negative, especially for CBC training. Overtime is positively related to quality. Experienced workers receive overtime and they make fewer mistakes than the overall workforce. Oakland's workload declined slightly over the last five years.

Pearl Harbor's overtime is inversely related to productivity. Workers may be less efficient when working overtime and do not produce as much as they would have in a comparable period during normal working hours. Nothing else is significant for Pearl Harbor.

At Puget Sound CBC training negatively impacted quality and timeliness. Physical distribution training positively impacted quality and productivity. Productivity is correlated with timeliness, and Q & T. The inverse correlation between overtime and timeliness, and Q & T is very strong.

San Diego is the only supply center that does not show some kind of relationship with either measure of productivity as a dependent variable.

Summary

The premise of CBC is that training will improve work quality, timeliness, and productivity. Proper training improves worker skills and attitudes, resulting in a more desirable on-the-job behavior, resulting in improved accuracy, timeliness and productivity. The analysis shows that a statistically significant change in performance occured, but was probably not caused by CBC training. At Puget Sound physical distribution training appears to have contributed to an improvement in performance and productivity.

The analysis presented in this paper does support some trends. Overall:

1. CBC training may not be related to performance;
2. physical distribution training is positively related to timeliness, Q & T, and receipt and issue productivity;
3. overtime is inversely related to quality.

Results differ for each supply center but nowhere does training appear to have significantly impacted warehouse worker performance.

At Jacksonville physical distribution training and CBC may be related to productivity. But, Jacksonville's workload increased while the number of workers remained the same resulting in an increased output per worker.

At Norfolk CBC may be positively related to productivity. Norfolk's workload declined while the number of workers remained stable resulting in an increase in productivity per worker.

At Oakland training is negatively related to all performance and overtime may be positively related to quality. Pearl Harbor and Charleston show no relationships at all with training.

SYSTEMATIC CURRICULUM REVIEW

Dr. Alexander M. Gottesman

Naval Health Sciences Education and Training Command

Trainers have been exceptionally successful in documenting how training is organized, conducted, and evaluated. Available for inspection or dissemination are clearly written curriculum outlines, programs of instruction, or other forms of course descriptions. At a minimum, course documentation identifies unit and lesson topic titles and specifies contact hours for didactic and practical training. Additionally, better course documents contain an overall goal statement or training aim as well as terminal and enabling learning objectives. In ideal course documentation the learning objectives are finely tuned to be consistent with tasks selected for training, presentation of instruction, and evaluation of task mastery; to be adequate in terms of intended final outcomes, measurability, and conditions under which performance is conducted.

CYCLICAL CURRICULUM REVIEW

Despite the excellence achieved in documenting the organization, presentation, and assessment of training, the assurance that training satisfactorily meets training requirements is rarely obtained. One workable system is underway at the Naval Health Sciences Education and Training Command (HSETC), Bethesda, Maryland. A Cyclical Curriculum Review (CCR) project based on the Instructional Systems Development Model (Gottesman, 1981) was begun in 1981. The purpose is to systematically review each of the 47 specific training programs leading to medical or dental specialties for Navy technicians. Table 1 summarizes the input, process, and product for each phase or step of the CCR.

Table 1 — Cyclical Curriculum Review

PHASE STEP	INPUT	PROCESS	PRODUCT
Analyze job	Curriculum outlines Task studies Evaluation reports Job reviews Professional guidelines	Develop job task inventory Validate job task inventory Conduct task training survey Analyze task training survey data	Tasks with high potential for training Tasks with low potential for training Tasks with ambiguous data Tasks added by responders Task training effectiveness index Task learning difficulty index
Select tasks for training	Tasks with high potential for training Tasks with low potential for training Tasks with ambiguous data Tasks added by responders	Convene jury of experts Submit task lists for final review by jury Resolve priority for training Analyze input from jury Prepare task lists	List of tasks selected for training List of tasks not selected for training
Review existing course	Tasks selected for training Current curriculum outline	Match selected tasks to objectives Develop planning document	Objectives to be deleted Objectives to be revised Tasks to be added to training Planning document for type revision
Design	Task learning difficulty index Training effectiveness index Objectives to be deleted Objectives to be revised Selected tasks to be added Planning document for type of revision	Develop revise objectives Develop student assessment procedures for new or revised objectives Determine sequence and structure	Revised curriculum outline Student assessment procedures and materials
Develop	Revised curriculum outline Previous training materials	Specify learning strategy Calculate and request required resources Review existing materials Revise and validate instruction	Revised instructor and student materials Validation report Course approval letter
Implement Control	Revised curriculum outline Revised instructor and student materials Course approval letter	Implement revised curriculum Implement evaluation strategy Implement changes resulting from evaluation	Revised curriculum implemented Record of minor curriculum changes Completed Cyclical Curriculum Review

BEST COPY AVAILABLE

JOB ANALYSIS

The CCR begins with a job analysis to compile an initial job/task inventory. Numerous basic sources are used to gather task statements or information that leads to the formulation of task statements. These include curriculum outlines for Navy medical or dental specialists' courses, occupational studies based on observed job analyses and surveys of tasks performed by the Naval Military Personnel Command, reviews of enlisted training submitted by Navy subject matter experts (SMEs) , medical technician training guidelines prepared by learned societies or professional organizations, and appropriate instructional materials from local, state, or federal agencies. The collected data from these sources are examined and a job/task inventory is developed by dental facilities. Task performance needs may differ depending upon the assignment location of a specific technician. Sea or shore duty; large, medium, or small duty station; and mission or function of the Navy facility will account for differences in task performance needs. But formal school training must be designed to meet needs Navy-wide--regardless of where any particular technician is located. Thus the validation process is necessary for further refinement of the initial job/task inventory. Tasks may therefore be added by the validation jury; task description and clustering changes may be recommended.

The validation phase is composed of the following six steps:

- Names and locations of SMEs are requested from the Commander, Naval Medical Command.
- SMEs are notified of their selection and the CCR purposes and procedures are explained to them.
- A job/task inventory response form is developed to ascertain whether technicians should or should not perform each of the tasks and to provide for the listing of additional tasks, as recommended by the experts.
- Responses from SMEs are collected through interviews or mailed response forms.
- Response data are analyzed for resolved discrepancies and any problems associated with clustering as well as determining what tasks should be added or deleted.
- A revised job/task inventory is prepared.

TASK TRAINING SURVEY

The validated job/task inventory is the major input for conducting a formal task training survey. NSEFC is provided extensive services by the Navy Occupational Development and Analysis Center (NODAC). Cooperatively, personnel from both agencies design the task training survey instruments. Specific directions for responding to a mark sense answer booklet are written to accompany the validated job/task inventory. One design is sent to either the entire community of specific Navy medical or dental technicians, or to a sample population of 100 to 200 if the specialty population exceeds 200 members. A second design is sent to the commissioned officers who utilize or direct the services of the technicians whose formal school curriculum is being studied.

Technicians participating in the task training survey supply demographic data and are asked to respond to two questions for each task statement:

- "Do you perform the task?"
- "If yes, how difficult was it to learn this task in relation to the other tasks you mastered?"

The second question is answered by marking a 1 to 8 scale for response options ranging from extremely low (difficulty) to extremely high (difficulty).

Commissioned officers are asked to respond to three questions in addition to providing data on their medical or dental specialty and the type and location of the facility to which they are presently assigned. Questions for commissioned officers are as follows:

- "Should the technician perform this task?"
- "If yes, how well were technicians prepared to perform this task prior to job entry?" [Responses are made for extremely low (preparation level) to extremely high (preparation level) on a scale ranging from 1 to 8.]
- "At what level is training recommended--familiarization only or hands-on/thorough knowledge?"

For the last question, the following explanation is provided (NODAC, 1982):

Familiarization means information should be provided on basic facts, components, capabilities, etc.

Hands-On/Thorough Knowledge means training should be provided which includes familiarization, but goes further with actual or simulated hands-on practice or in-depth knowledge requiring judgment or application of theory.

Responses from the enlisted personnel provide, by total and by subgroups, the percentage of technicians performing each task and data from which a task difficulty index can be derived. Data from the commissioned officers provide, by total and by subgroups:

- Percentages of responders who believe a task should or should not be performed.
- Information from which a task effectiveness index can be derived.
- The level to which each task should be trained (familiarization only or hands-on/thorough knowledge).

ANALYSIS OF TASK TRAINING SURVEY DATA

Computer printouts on task training survey response data prepared by NODAC are carefully reviewed and summarized by HSETC education specialists. Each task training survey solicits comments and suggested additional tasks from responders. These are also reviewed and summarized. Specific criteria are applied to these summaries to determine whether tasks have a high or low potential for formal school training and to classify selected tasks according to training priority. A task is considered having a high potential for training if 50 percent or more of the commissioned officers indicate that technicians should perform the task or if 50 percent or more of the technicians indicate that they do perform the task. Additionally, if a majority of any subgroup respond positively, the task is included in the high-potential group. Low potential for training is assigned to a task when less than 40 percent of the responders indicate that the task is or should be performed.

A conservative approach is taken when priority for training is determined. If 40 percent or more of the total sample of commissioned officers indicate a task should be trained to the hands-on/thorough knowledge level, that task is assigned the higher priority training level.

Training effectiveness and task learning difficulty indices based on task training survey response data are constructed and provided to formal schools personnel. Table 2 is an example of a report of training effectiveness and learning difficulty of tasks selected for training in the Cardiopulmonary Technician program.

RESOLVING AMBIGUITIES

Primarily, the analysis of response data yields two task lists:

- Tasks with a high potential for formal school training.
- Tasks with a low potential for formal school training.

If the response data are inconclusive regarding whether tasks should or should not be selected for training, such as above 40 percent from both the commissioned officers and the technicians or above 40 percent in two or more subgroups, a third list of tasks is prepared. A fourth list consisting of tasks added by responders is also prepared. To obtain two lists only, that is, tasks selected for formal school training and tasks not selected for formal school training, the four lists are sent to the jury of experts for final resolution. The jury is instructed to:

- Review all lists.
- Place tasks from list three (ambiguous data) into list one or two.
- Place appropriate tasks from list four into list one.
- Indicate the level of training (familiarization only or hands-on/thorough knowledge).

Table 2
Training Effectiveness/Learning Difficulty Of Tasks Recommended For Training Cardiopulmonary Technicians

TRAINING EFFECTIVENESS INDEX

	Expected Mean Ranges	Actual Mean Ranges
1 Highly effective	5.51 - 8.00	5.53 - 7.00 (152)*
2 Effective	4.01 - 5.50	4.82 - 5.50 (54)
3 Least effective	1.00 - 4.00	3.67 - 3.92 (2)

The Training Effectiveness Index shows the relative effectiveness of training as perceived by commissioned officers responding to the survey. Tasks rated highly effective on the index were seen as more effectively trained in the formal school than were tasks rated less effective. This does not necessarily mean that training for tasks rated "Least effective" was actually ineffective.

LEARNING DIFFICULTY INDEX

	Expected Mean Ranges	Actual Mean Ranges
1 Most difficult	5.51 - 8.00	5.52 - 5.93 (4)*
2 Moderately difficult	4.01 - 5.50	4.03 - 5.50 (74)
3 Least difficult	1.00 - 4.00	2.03 - 4.00 (130)

The Learning Difficulty Index shows the relative difficulty of learning the tasks, as perceived by the technicians responding to the survey. As with the effectiveness ratings, these are comparative only.

The rating scale on the survey was as follows:

4.00 Below average	5.00 Above average
2.00 Very low	6.00 High
3.00 Low	7.00 Very high
4.00 Below average	8.00 Extremely high

* number of tasks in range is shown in ( )

REVIEWING EXISTING COURSE

Conduct of the previous activities--draft job/task inventory, job/task inventory validation, task training survey, task training survey response analysis, and ambiguous data resolution--results in the following products:

- Tasks selected for formal school training list.
- Tasks not selected for formal school training list.
- Training effectiveness index.
- Task learning difficulty index.

The formal schools offering the specific curriculum under review receive copies of the two lists and indices and are briefed on the CCR process. With review and assistance of USAFC, program directors and lead instructors at the formal schools are requested to use the task lists to accomplish the following:

- Identify learning objectives that adequately support tasks selected for formal school training.
- Identify learning objectives that support tasks selected for training but require revision.
- Eliminate learning objectives that do not support tasks selected for formal school training.
- Identify selected tasks that are not supported by learning objectives.
- Cross-index adequate learning objectives with tasks selected for formal school training.

The number and complexity of tasks for which training is not currently provided and/or the number of learning objectives that require development

BEST COPY AVAILABLE

or revision determine the type of curriculum revision mandated by CCR. Three types of revisions have been identified as follows (CNET, 1981):

- Type A--Changes in course length, objectives, and subject matter to such an extent that logistic support, personnel allocations, funds, and the like are affected. Require project plans and HSETC approval.
- Type B--Modifications within the established structure of the course, including major rescheduling of topics, time, or revision of instructional procedures. Require a Plan of Action and Milestones developed in conjunction with the formal school(s).
- Type C--Minor changes such as correction of clerical errors; insertion of titles or designations of films, publications, and equipment; minor adjustments in time allocations; and addition of learning activities. Require HSETC executive correspondence setting forth minor course changes and tentative completion dates.

REVISION PLANNING DOCUMENT

Cooperatively, the formal schools and HSETC determine the resource constraints that will affect effective and timely completion of the curriculum revision. The type of revision (A, B, or C), the number and experience level of instructors for the program under review, whether the program is offered at a single or multiple training site(s) and the type and number of curriculum support personnel at the training site(s), and the number of programs in the various stages of CCR process help determine resource constraints. This determination affects the level of responsibility assigned to the formal school(s) or assumed by HSETC for the design and development phases. HSETC education specialists and school personnel estimate completion dates for these two phases.

DESIGN PHASE

At the design phase, which is actually a redesign of the curriculum, HSETC and the schools consider task learning difficulty and training effectiveness data resulting from the analysis phase. The course review conducted prior to this phase provides a list of learning objectives to be deleted from the course and a list of learning objectives to be revised. Also provided is a list of selected tasks to be added to the course. After deleting learning objectives not supportive of tasks selected for training, the formal schools, with assistance from HSETC develop newly required objectives. This process includes revising existing learning objectives, where appropriate, or writing new learning objectives for selected tasks not currently provided for in the curriculum. Student assessment procedures are developed for all new or revised learning objectives. Both performance checklists and criterion-referenced test items are created or redesigned for evaluation of students. Finally, the sequence and structure of the entire curriculum are determined. These are tentative decisions on the time required for training students to master the tasks selected for training and the point or place in the training program where each learning objective will occur. Necessary adjustments in sequence and structure may follow a pilot phase of the revised curriculum.

DEVELOP PHASE

By reviewing appropriate previously used training materials and the revised curriculum outline, the formal schools with HSETC assistance:

- Specify learning strategies.

- Calculate resource requirements of time, manpower, and cost to implement revisions, and request resource approval from higher authority if Type A revision is required.
- Review existing materials.
- Develop revised instruction.
- Validate revised instruction.

As a result, program directors and instructors of the training program under review will have revised lesson topic guides and supporting materials available for students. These documents will be used in the pilot phase to validate instruction. A validation report is then submitted by the formal schools and a course approval letter is provided by HSETC. If subsequent additions or changes are deemed necessary, higher authority action on a request for resource allocation is obtained.

IMPLEMENT/CONTROL PHASES

The final two phases are combined to implement the revisions and evaluate the revised curriculum. The formal schools conduct the training of students using any adjustments to the curriculum found necessary during the pilot or validation phases of instruction, while HSETC monitors the outcome. The evaluation or control phase strategy uses a course evaluation plan with course review checklists. All evaluative data are analyzed and necessary changes are identified. If changes are required as a result of the evaluation summary report, the schools submit to HSETC a request to approve minor curriculum changes. At this point, the first cycle of the curriculum review is completed.

SUBSEQUENT CYCLES

Subsequent cycles will occur but will require far less time because of the work accomplished during the first cycle. Specifically, tasks lists requiring only minor revisions will be available. Learning objectives will be task based, and only new objectives necessary to support additional tasks selected for formal school training during subsequent cycles will need to be developed. Further, the ongoing evaluation plan and a minor curriculum changes procedure will keep each curriculum up-to-date. Finally, after the initial cycle, less critical specialty courses may be scheduled for a CCR on alternate cycles rather than every cycle.

SUMMARY

The Cyclical Curriculum Review procedure is based on a systems approach to curriculum development. It uses subject matter as well as process experts to select tasks for training based on state-of-the-art content that will produce highly competent medical and dental technicians. Through the CCR process, the effectiveness of these technicians in meeting the needs of the Navy is ensured.

REFERENCES

1. Gottesman, A.M. (1981, April). "Instructional Systems Development." Journal of Systems Management, 6-9.
2. Navy Occupational Development and Analysis Center (1982, May). Navy Occupational Task Analysis Program. Washington, D.C.: Navy Military Personnel Command.
3. Chief of Naval Education and Training (1981, September). Procedures for Instructional Systems Development. Pensacola, FL: Naval Education and Training Command.

Computer Aversion as a Source of Bias in Computerized Testing

Jo Anna Wood and Gordon F. Pitz
Southern Illinois University, Carbondale

Tests, conventional and computerized, are a pervasive aspect of our lives. They are used as part of placement and selection procedures for employment, training and educational opportunities. In all these instances, test scores are used to classify individuals on the basis of ability or aptitude, as measured by the test. When the testing procedure itself interferes with the measurement of relevant phenomena in a non-random way, that procedure will discriminate against certain test-takers, by incorporating an irrelevant attribute into the classification process.

A growing body of literature (e.g., Lawton & Gerschner, 1982; Naiman, 1982; Nickerson, 1981) suggests that a person's beliefs and attitudes about human-computer interactions will affect that person's ability to interact with a computer. The term computer aversion has been adopted by Meier (1984) to refer to such negative beliefs and attitudes. Computer aversion can hamper one's performance on computerized tasks such as data input, word-processing, or using databases. It seems likely that similar effects would be found when the task is a computerized test.

The research presented in this paper is an attempt to address systematically some shortcomings of published findings on computerized testing and computer aversion, as well to provide some validity data for a measure of computer aversion. The hypotheses to be tested concerned two dependent variables, state anxiety and test performance (number of errors). It was expected that computer aversion would negatively bias the results obtained in computerized testing. Further bias was expected when the testing program was "unfriendly" or difficult to use. In addition, this research was designed to determine whether computer aversion is different from two possibly related concepts, test anxiety and trait anxiety.

Methods

Subjects

Subjects (N=92) were recruited from non-college student adult populations in the Alton and Carbondale, Illinois areas. Subjects were recruited from populations of hospital in the two areas. Only English-speaking subjects were recruited to avoid confounding test results with language abilities.

Tests and other measures

Verbal and math questions were selected from published SAT tests (College Entrance Examination Board, 1983). Two tests were constructed, each composed of two math and three verbal subtests. The tests used a multiple choice format, and were timed. Each test was developed for both paper-and-pencil and computer administration.

Meier (personal communication, January 5, 1985) has developed a questionnaire to measure computer aversion among clinical psychologists (Computer Attitudes Scale). This scale was modified for use with health care personnel for use in the present study. The State-Trait Anxiety Inventory was used to obtain measures indicative of participants' levels of anxiety during testing (state anxiety),

BEST COPY AVAILABLE

as well as to obtain a measure of general (trait) anxiety. The Achievement Anxiety Test (Alpert & Haber, 1960) was designed to measure subjects' perceptions of the extent to which anxiety is either facilitating or debilitating of test performance.

Computerized test administration and scoring were handled by computer programs written in Pascal and developed for this purpose. Computerized tests were administered on a 16-bit IBM-PC compatible machine.

Procedures

Subjects were randomly assigned to either a friendly or unfriendly program condition. Total allotted time was the same for both conditions. The unfriendly program forced subjects to work through each item at a fixed pace and required complex responses. These demands were expected to cause many errors. The friendly program allowed subjects to allocate their own time within each subtest, and used simple response procedures. Subjects were further randomly assigned to receive either the paper-and-pencil or computerized test first. Versions of the abilities tests were randomly assigned as either paper-and-pencil or computerized administration for each subject. Random assignment was handled through a computer program, with the only restriction being an equal number of subjects in all conditions.

Results

Analysis of Covariance (ANCOVA) was used to test the effects of program friendliness and test mode (i.e., computerized vs paper and pencil) on performance and state anxiety while controlling for the effects of computer aversion. An additional factor, order, was included to determine if the dependent measures were influenced by the order in which the test modes were used.

For verbal performance the significant effects of interest are interactions of Program (friendly vs. unfriendly) by Mode (computer vs. paper) ($F(1,84) = 8.69$, $p=.004$), and Order (computer first vs paper first) by Computer Aversion by Mode ($F(1,84) = 5.65$, $p=.020$). The former effect was also significant in the analysis of overall performance. The latter was not; overall performance included scores on math subtests that probably represented mostly random error.

The two-way interaction of program by mode is shown in Table 1. Both forms of the computer programs induced poorer performance than did the paper and pencil tests, but the effect was greater for the "unfriendly" version.

The significant three-way interaction indicates that one or more of the cells in the design matrix differed from others in terms of the relationship between computer aversion and performance. Separate regressions of performance on computer aversion were calculated for each cell. As shown in Table 2, computer aversion was significantly and positively related to performance on computerized tests only when those tests preceded the paper and pencil tests. Computer aversion was not related to performance on paper tests.

When test anxiety and trait anxiety were introduced into the within-cells regression models, the significant regression coefficients for computerized testing remained high. In general, adding the extra covariates had little effect on size of regression coefficients for computer aversion.

Multiple Regression was used to examine further the relationship between computer aversion and test and trait anxieties. Neither test anxiety nor trait anxiety proved to be a good predictor of computer aversion (R^2=.135, F (2,89) = .81998, p = .4437).

State anxiety was affected by two significant sources of variation : Mode, and a 3-way interaction of Order, Program and Computer Aversion. Subjects were significantly more anxious after computerized testing than after paper testing (F (1,84) = 13.91, p=.000). The significant three-way interaction (F (1,84)= 4.74, p=.032) indicates that the relationship between computer aversion and state anxiety differed across cells in the design. Separate regressions were calculated for each cell; only the regression coefficient for paper testing following "friendly" computerized testing was significant (B=.55).

Discussion

The first important finding of this study is that Computer Aversion is not just another form of anxiety. It is a separate construct unrelated to the propensity toward anxiety or fear of testing.

Second, computerized testing provoked more anxiety than paper testing, especially when the program was unfriendly. Interestingly, the reported state anxiety during computerized testing was not significantly related to computer aversion. People were more anxious during computer testing, but this had nothing to do with their previously expressed attitudes about interacting with the computer. It appears that using a computer, and the difficulty involved in using the software, are sources of stress unrelated to one's expectations about interacting with computers, and thus could be termed "computer anxiety."

Third, and most importantly, this study demonstrates that computerized tests are not necessarily equivalent to paper-and-pencil tests, even when the contents are identical. Apart from generally poorer performance, changing the test mode (i.e., using the computer rather than a paper test) introduced a new source of variation among scores, namely, computer aversion. A person's beliefs and feelings about interacting with a computer can affect that person's interactions with computers. In this study it affected people's ability to demonstrate their cognitive abilities. In other words, computerized test scores were seriously biased against those people who were computer aversive.

Figures 1 through 3 illustrate the magnitude of bias against computer aversion for subjects experiencing "unfriendly" and "friendly" computer tests preceding paper tests. Figure 1 represents the proportions of high computer aversives and low computer aversives whose scores fall at or above each score on the paper test. The median score for both groups combined (grand median) was 19 points. In each group, 57% of the subjects had scores at or above the grand median. On the "unfriendly" computerized test, 75% of low computer aversives had scores at or above the grand median, while only 36% of the high computer aversives exceeded the median. The difference was even more pronounced on the "friendly "

Table 1

Cell means and Standard Deviations for Math and Verbal Errors

		Friendly Program		Unfriendly Program	
COMPUTER TEST		Mean	S.D.	Mean	S.D.
Computer First	Math	14.48	(2.89)	16.74	(2.30)
	Verbal	14.43	(5.04)	16.26	(6.80)
Paper First	Math	15.13	(3.68)	16.26	(2.65)
	Verbal	12.87	(4.59)	19.22	(5.48)
PAPER TEST					
Computer First	Math	13.17	(3.76)	13.78	(4.06)
	Verbal	12.43	(5.67)	12.04	(6.67)
Paper First [1]	Math	14.48	(4.24)	14.48	(4.24)
	Verbal	13.83	(6.40)	13.83	(7.69)

[1]Cell means for these two conditions were calculated as a single group, since the two conditions were equivalent.

Table 2

Regression Coefficients for Verbal Errors on Computer Aversion

	Friendly		Unfriendly	
	b	beta	b	beta
COMPUTER TEST				
Computer First	.31	.49*	.52	.54*
Paper First	.08	.09	.08	.15
PAPER TEST				
Computer First	.20	.29	.17	.18
Paper First[1]	.17	.21	.17	.21

NOTES: (1) Starred items are significant at $p<.01$.
(2) b's are raw regression coefficients, while Beta's are standardized regression coefficients.

[1]These two groups were treated as single cell for regression

computerized test. 82% of the low computer aversive, but only 25% of the high computer aversives, had scores at or above the grand median.

The last comparison is most interesting because the "friendly" computer test most closely approximates computerizd testing procedures that would be used in selection and classroom testing applications. Although it has less overall effect on performance, this test appeared to be more biased than the "unfriendly" version. Insofar as this experiment represents computerized testing for selection purposes, it appears that computerized tests have a built-in bias against computer aversives.

If the results of the present study are indicative of what occurs in other computerized testing applications, then one of two things may occur. First, Computer Aversion may be related both to computerized test performance and to performance on some criterion measure (e.g., job performance, success in college). In this case, using computerized tests should result in more accurate predictions of the criterion. In the second scenario, Computer Aversion is related to computerized test performance, but not to performance on the criterion measure. If this is true, then Computer Aversion acts as a moderator variable; that is, it affects the relationship between computerized test performance and performance on the criterion, and does so differently for various groups of subjects.

References

Alpert, R. & Haber, R. (1960). Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 61, 207-215.

College Entrance Examination Board (1983). 10 SATs Scholastic Aptitude Tests of the College Board, New York:CEEB.

Lawton, J. & Gerschner, V.T. (1982). A review of the literature on attitudes towards computers and computerized instruction. Journal of Research and Development in Education, 16, 50-55.

Meier, S.T. (1984). Computer Aversion. Unpublished manuscript, Missouri Institute of Psychiatry, St. Louis. MO.

Naiman, A. (1982). Women technophobia and computers. Classroom Computer News, 2, 23-24.

Nickerson, R. (1981). Why interactive computer systems are sometimes not used by people who might benefit from them. International Journal of Man-Machine Studies, 15. 469-483.

FIGURE 1: PAPER TEST SCORES
□ = Hi averse; △ = Low averse

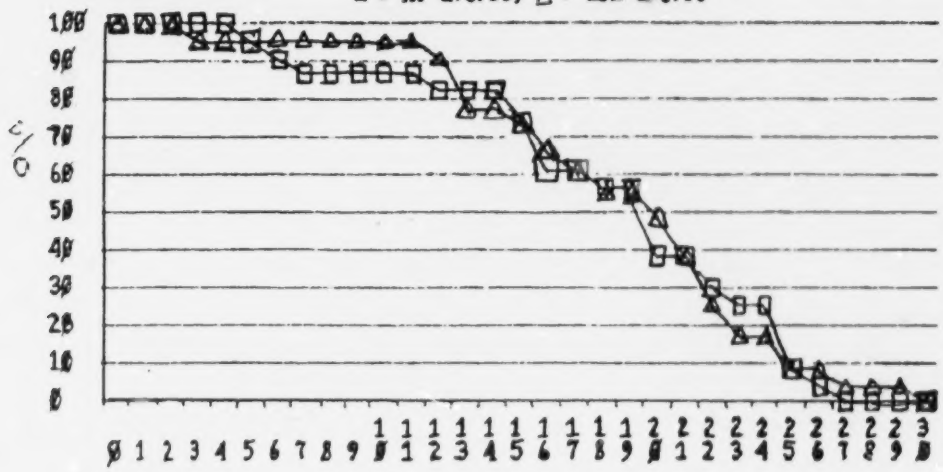


FIGURE 2: UNFRIENDLY COMP. TEST SCORES
□ = Hi averse; △ = Low averse

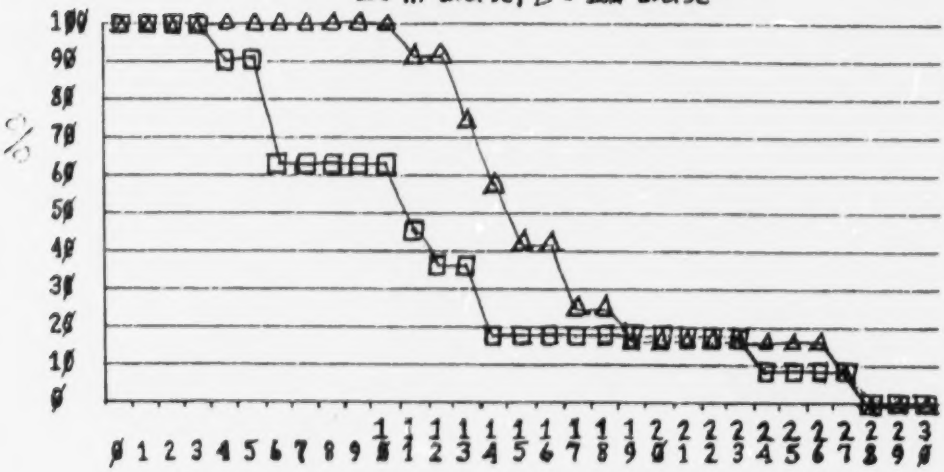
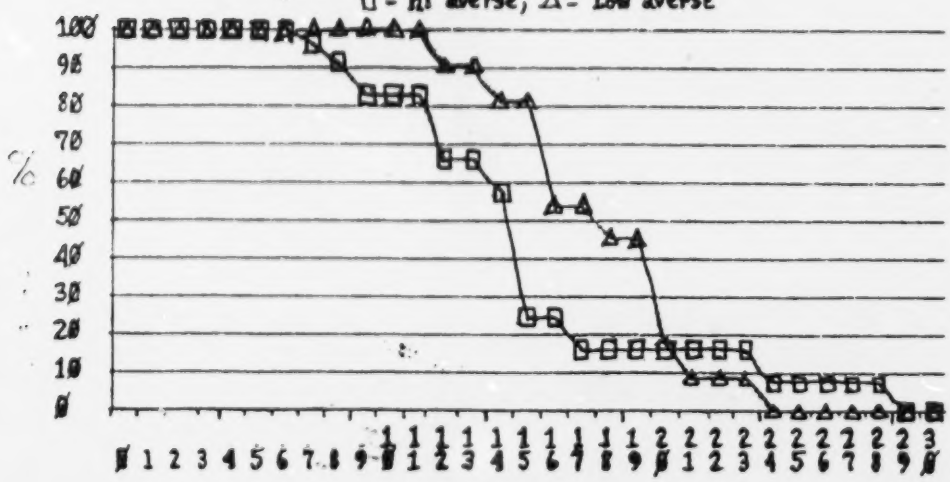


FIGURE 3: FRIENDLY COMP. TEST SCORES
□ = Hi averse; △ = Low averse



Orientation, surrogate travel, and gender differences in videogame strategy

Sharon Tkacz
Army Research Institute

Research has shown that individuals vary in their ability to process and use spatial information (Kosslyn, Brunn, Cave, & Wallach, 1983). They also differ in the frame of reference they use to form memory representations of space (Sholl & Egeth, 1980). Some use an egocentric system (e.g., right or left) whereas others use a topographic system (e.g., north or south). Developmental studies (Pick & Reiser, 1982) have suggested that individuals adopt the topographic, more sophisticated system as they mature.

Piaget has suggested that physical movement through the environment is how spatial reasoning skills are acquired. Goldin and Thorndyke (1981) support Piaget's arguments, demonstrating that navigation through space provides a unique kind of spatial knowledge, procedural knowledge, that cannot be acquired simply by reading maps.

The fact that movement leads to procedural knowledge acquisition has been applied to navigation training. Cohen (1980) has shown that the information derived from actual travel can be approximated by surrogate travel. In some cases, simulated movement may be even more effective as a training aid than actual navigation: if only relevant information is presented, irrelevant information cannot be distracting.

Much research (Wittig & Petersen, 1979) has demonstrated gender differences in spatial information processing, including the relationship of cognitive variables to sex-role identity. Since previous research has shown that individuals tend to conform to sex-role expectations, and expertise in computers and videogames is considered masculine, it is reasonable to predict that females may not perform as well as males. Whether lower female performance is due to a lack of cognitive ability or adherence to sex-role expectations has not been established.

The research described below investigated navigation through an artificial environment created by a microcomputer. The game required players to use a topographic reference system to indicate directions. In addition to dependent measures derived from the videogame, cognitive components assumed to underly game performance were assessed. Cognitive components were also examined to see if females and males differ in the basic cognitive skills required by the game.

Method

One hundred and ninety undergraduates served as participants. They were administered several psychometric tests, and played a series of eight videogames requiring them to escape from a 5 x 5 x 5 cubic maze. Players moved from one room to the next through openings in the floors, walls, and ceilings by typing "n", "s", "e", "w", "u", or "d" for north, south, east, west, up or down directions. Two types of information available

were current Position (P) and location of the escape room or Goal (G), each defined by x, y, z coordinates. Combinations of these two types of information formed the four INFORMATION CONDITIONS: PG, P, G, or Q.

In the PG condition, information on both the player's Position (P) and the Goal (G) was provided continuously on the screen. Subjects in the P or G condition had only one kind of information available. Those in the Q condition had no information displayed but were permitted to request both P and G coordinates. All participants played four PG games, in order to familiarize them with the game and the keyboard, and then four more games, in one of the four INFORMATION CONDITIONS.

Results and Discussion

Males showed superior spatial performance. (Table 1 shows means for all psychometric tests.) There were no gender differences on vocabulary or reasoning tests. These data indicate that males and females differ significantly in the skills they bring to the experiment that were expected to underly game performance.

Table 1. Gender differences in psychometric measures.

Variable	Female Mean	Male Mean	t	p
Abstract Orientation	101	113	3.4	.001
Map Orientation	8.5	10.2	2.7	.01
Figural Reasoning	22.6	22.4	.2	*
Mental Rotation	25.6	27.7	2.2	.05
Vocabulary	55.3	55.2	0	*

Game performance was described by ten dependent variables derived from individual key presses for each game. SCORE indicates the total time to complete one game. RESPONSE TIME indicates mean time between any two keypresses. Similarly, STATIONARY TIME indicates mean time spent in a room. EFFICIENCY is a ratio of the minimum distance to actual distance between starting Position and Goal room. REORIENTATION is the rate (the number of times per minute) that players changed the direction they were facing. SURFACE RATE is a measure of time spent in surface rooms of the cube. Similarly, INTERIOR RATE is a measure of time spent in interior rooms. VISIBLE CRASH indicates how many times a player tried to go through a wall (visible on the screen) that did not have a door. Similarly, REAR CRASH indicates how many times a player tried to go through the wall behind them (not visible on the screen). Lastly, ERROR KEY is a measure of illegitimate key presses.

An analysis of variance was performed on the ten dependent measures with two between-subjects variables (INFORMATION CONDITION, C and GENDER, G) and one within-subjects variable (PRACTICE, P). SCORE and REORIENTATION were the only dependent measures for which any GENDER (G) effect was obtained (see Table

Table 2. Analyses of variance on ten videogame measures.

Dependent measure	Source of variation	F	p
Score	C	76.3	.001
	G	11.6	.001
	P	15.3	.001
	CP	3.0	.001
	GP	2.9	.05
Response time	P	55.3	.001
Stationary time	C	18.1	.001
	P	30.0	.001
	CP	3.7	.001
Efficiency	C	64.8	.001
	P	2.9	.05
Reorientation	G	7.1	.01
Surface rate	C	6.7	.001
	P	40.9	.001
	CP	2.7	.01
Interior rate	C	27.2	.001
	CP	1.9	.05
Visible crash	C	11.6	.001
	P	9.6	.001
	G	4.1	.05
Rear crash	C	51.3	.001
	P	8.5	.001
	CP	5.1	.001
Error key	C	9.2	.001
	P	3.8	.01
	CP	2.5	.01

2). Although almost all measures improved with PRACTICE (P), the GENDER x PRACTICE interaction (GP) was significant only for SCORE, suggesting that, while their initial scores may be lower, females may show greater improvement. The effect of INFORMATION CONDITION (C) was significant for all variables except RESPONSE TIME and REORIENTATION, indicating that the rate of key pressing and the rate of turning is independent of the amount and type of information available. Finally, a PRACTICE x INFORMATION CONDITION interaction (CP) was obtained for several measures - SCORE, STATIONARY TIME, SURFACE RATE, INTERIOR RATE, REAR CRASH, and ERROR KEY. The fact that this interaction was not obtained for all dependent variables indicates that improvement in

performance is not simply a function of repeated practice. These game-derived performance measures may be indices of individual differences in information-processing capacity, and not subject to practice effects.

Means for these dependent measures are shown in Table 3. Here, SCORE and REAR CRASH were the only variables for which any gender difference was obtained, in only two INFORMATION CONDITIONS (G and Q). These results suggest performance is very similar for males and females.

Table 3. Results of t-tests on videogame measures.

Variable	INFORMATION CONDITION	Female Mean	Male Mean	t	p
Score	PG	82	61	1.197	*
	Q	142	99	1.656	*
	G	414	247	3.073	.01
	P	639	507	1.493	*
Response time	PG	4.4	4.0	.608	*
	Q	3.8	3.7	.182	*
	G	3.5	3.8	.651	*
	P	3.7	3.3	.987	*
Stationary time	PG	5.5	5.1	.457	*
	Q	9.8	8.8	.618	*
	G	4.7	4.8	.170	*
	P	5.2	4.6	1.075	*
Efficiency	PG	.60	.59	.212	*
	Q	.56	.63	1.246	*
	G	.21	.21	0	*
	P	.11	.11	0	*
Reorientation	PG	.07	1.16	1.726	*
	Q	.09	.47	1.212	*
	G	.10	.31	1.243	*
	P	.09	.20	.808	*
Surface rate	PG	8.28	8.52	.213	*
	Q	5.95	5.42	.581	*
	G	6.45	6.90	.623	*
	P	7.07	8.83	1.872	*
Interior rate	PG	3.59	3.68	-.127	*
	Q	1.65	2.26	1.954	*
	G	1.33	1.72	1.047	*
	P	1.01	1.05	.119	*
Visible crash	PG	1.72	1.49	.572	*
	Q	.85	.88	.103	*
	G	3.84	2.38	1.692	*
	P	3.02	2.10	1.530	*
Rear crash	PG	.69	.70	.037	*
	Q	.48	.21	2.808	.01
	G	1.36	1.28	.293	*
	P	2.38	2.47	.224	*
Error key	PG	.11	.20	.976	*
	Q	.40	.52	.661	*
	G	.08	.12	.679	*
	P	.13	.22	.584	*

Data in Table 2 also indicate that performance varies across INFORMATION CONDITION (C) for every variable except RESPONSE TIME, suggesting that different strategies are employed in different INFORMATION CONDITIONS. Rather than indicating different difficulty levels of the same game, INFORMATION CONDITIONS may be qualitatively different games, from a problem-solving perspective. That is, a task that involves finding a goal without knowledge of your own position may not have much in common with a situation where your position is known.

This interpretation is supported by results of stepwise multiple regression analyses. SCORE was predicted from the psychometric measures, shown in Table 1, for males and females separately, and for males and females combined, for each INFORMATION CONDITION. Table 4 shows amount of variance in SCORE accounted for by the best combination of two psychometric predictors. The different INFORMATION CONDITIONS have different psychometric predictors, suggesting differences in cognitive components.

In contrast to the absence of gender differences in Table 3, data in Table 4 indicate that components of performance differ for females and males. This difference is particulary clear for condition Q. Cognitive correlates of female performance are vocabulary and reasoning, neither of which are spatial measures. Conversely, the best predictors of male performance are mental rotation and abstract orientation. Taken together with the data in Table 1, these results suggest that individuals may develop strategies that depend on their own skills, rather than strategies that are task dependent.

Table 4. Multiple regression analyses: predicting SCORE from psychometric measures.

INFORMATION CONDITION		Predictors	R-SQUARED
PG	females	reasoning, vocabulary	.453
	males	reasoning, abstract orientation	.556
	both	reasoning, vocabulary	.360
P	females	map & abstract orientation	.174
	males	map & abstract orientation	.320
	both	map & abstract orientation	.211
G	females	map & abstract orientation	.254
	males	map & abstract orientation	.177
	both	map orientation, reasoning	.186
Q	females	reasoning, vocabulary	.206
	males	abstract orientation, mental rotation	.409
	both	map orientation, mental rotation	.177

Summary & Conclusions

Differences in INFORMATION CONDITIONS demonstrate that this variable represents different task requirements. Thus, the dependent variables selected to describe game performance seem to do so adequately, since they reflect different aspects of the players' performance and strategy.

Regression analyses indicate that cognitive components underlying game performance are not the same for males and females. Although components underlying videogame performance differ, suggesting gender-related strategy differences, actual game performance shows little variation attributable to gender.

In sum, given that videogame skills are well retained, fun and relatively easy to acquire, they have much potential as instructional tools. For example, games simulating navigation could provide a simple, cost-effective way of training spatial learning strategies and exercising navigational skills. Since individuals that have different cognitive skills demonstrated similar game performance, different strategies may be employed to achieve the same results. Future instructional paradigms should provide for flexibility in strategy development so that learners may make the best of their individual cognitive strengths. Further, the dependent measures employed here reflect complex, strategic behavior in a simulated environment. Measures such as these may have greater ecological validity than standard psychometric tests in predicting individual differences in complex, real world behavior.

References

Cohen, M. E. (1980). The effects of environmental interactions on the structure and process of cognitive mapping. Ph.D. Thesis, Temple University, Philadelphia.

Goldin, S. & Thorndyke, P. (1981). Spatial learning and reasoning skill. Santa Monica, CA: The Rand Corporation, R-2085-ARMY.

Jones, M. (1984). Videogames as psychological tests. Simulation and Gaming, 15(2), 131-157.

Kosslyn, S. M., Brunn, J. L., Cave, C. R., & Wallach, R. W. (1983) Components of mental representation, (Contract No. N00014-79-C-0982) Arlington, Virginia: Office of Naval Research.

Pick, H. L., & Rieser, J. J. (1982). Children's cognitive mapping. In M. Potegal (Ed.), Spatial abilities: Development and physiological foundations (pp. 107-128). New York: Academic Press.

Sholl, M.J., & Egeth, H.E. (1980). Interpreting direction from graphic displays: Spatial frames of reference. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), Processing of Visible Language (Vol. 2). New York: Plenum Press.

Wittig, M. A., & Petersen, A. C. (Eds.). (1979). Sex-related differences in cognitive functioning. New York: Academic Press.

The Navy's General Unrestricted Line Community: Career Management and Career Development Problems

Gerry Wilcove

Navy Personnel Research and Development Center
San Diego, CA 92152

Community Description and Background

In 1981, General Unrestricted Line Officers (i.e., General URLs) were given community status, and a community manager was selected. Given its short history, it is not surprising that the community is a relatively unknown entity to people both within and outside the Navy. The community is composed of approximately 3,000 officers, 80-percent of whom are women. A disproportionate number of officers are lieutenant or below, although it is expected that the number of officers selected for executive officer, the second in command, will triple from 1984 to 1987. The function of this community is to support the Navy's fighting forces by serving in shore billets as general managers and specialists in areas such as personnel management, financial administration, data processing and computer technology, organizational effectiveness, and communications. There are also limited numbers of General URLs in operations systems technology, naval systems engineering, political-military strategic planning, and weapons engineering. Women are eligible to receive some training on combat vessels and may even become surface warfare officers. However, duty on combat vessels is restricted by law to temporary duty under noncombat conditions.

Career Management and Career Development Problems

I will first discuss some of the problems that have made it difficult for the General URL community to manage the careers of its officers and for individual officers to develop their careers. I will then describe the policy and procedural changes that occurred in 1984 that ameliorated some of the problems. The problems that are discusssed were identified from: (1) 1982 questionnaire results from approximately 45-percent of the community, (2) open-ended comments from the same questionnaires (N=500), (3) 50 interviews with community members, and (4) conversations with Washington managers and policy-makers. The problems are not listed in any kind of rank order.

First, in the policy-making area, the community has been handicapped by the lack of freedom to control the yearly number of accessions. Instead, these numbers have been determined for the community in accordance with a "surge-tank" concept; i.e., in accordance with estimates of the number of billets, or assignments, that would be vacant X number of years in the future because of a lack of warfare specialists. Managing careers becomes difficult when officers from different commissioning years are subjected to different selection ratios at the

same points in their careers. By the same token, a woman developing her career faces minimal or intense competition depending on her year of commissioning.

A second problem for General URLs has been the fact that they have been reassigned upon the completion of a tour by officers from another community; i.e., Surface Warfare Officers (SWOs). Thus, there has been less freedom to groom those of demonstrated potential for high responsibility senior jobs. Individual officers have complained that SWOs do not know the duties and requirements of existing assignments or their career potential.

A third career management and development problem has been the lack of high level acceptance concerning the General URL's right to career enhancing billets. This obstacle has been manifested in two ways: (1) the tendency of the Navy to assign warfare specialists, instead of General URLs, to career-enhancing shore billets, when a direct competition occurs, and (2) the tendency to reserve certain billets for warfare specialists, even though General URLs may be capable of performing the work. Here, career managers are stymied in their attempts to develop officers for major shore commands because of restricted opportunities at lower levels. Conversely, officers attempting to develop their careers in particular directions have been forced to reformulate their career goals.

A fourth problem has been a complaint about the quality of the billets available for General URLs. A LT perhaps best summarizes the feelings of dissatisfied General URLs when she states:

> "I feel that General URL jobs are overall the least desirable within the Navy. These jobs are poorly defined, usually not operational or competitive, and require no special education or background. Many are nonessential and have no clear career path associated with them. They can be filled by anything from a CWO2 to a LCDR and are often gapped for long periods of time" (i.e., left vacant).

A fifth problem is that the career path for General URLs has been an ambiguous one defined mainly in general terms; i.e., an individual should be sure to perform well in subspecialty and leadership positions. It has been difficult for career managers (i.e., the community manager, senior officers, assignment managers, and subspecialty managers) to offer advice when the "wickets" for career advancement were largely unknown, the criteria for promotion in competition with warfare officers were unformulated, and the career advancement potential of various subspecialties was a matter of conjecture.

This career advice problem showed up in survey results. For example, only 45-percent reported that they had been counseled on the "tickets" that have to be "punched" for career advancement, and only 25-percent said they had been counseled on the "blind alleys" that might destroy their careers.

The sixth problem concerns senior officers. The Navy has thus far concentrated on defining the career path for junior officers. Individuals who have completed a commander-commanding officer tour complain about the lack of a career pattern after that point and the lack of opportunities in the upper levels

of the Navy hierarchy. The latter includes policy-making positions in the "fifth wing of the Pentagon", senior administrative jobs at headquarter and staff commands (e.g., SURFPAC, RECRUITCOM, the Sixth Fleet), and commanding officer and executive officer billets at naval training commands and administrative commands. At the present time, there is no overall Pentagon group that formulates policy for the General URL community as there are for the other unrestricted line communities (SWOs, aviators, and submariners).

A seventh problem concerns training opportunities and pipelines, neither of which have been institutionalized. Survey results showed that only 25-percent had received training enroute to their new assignments. In addition, there is no department head school for General URLs, and General URLs do not attend the Prospective Executive Officer or Prospective Commanding Officer School unless they have been selected for major command.

The eighth problem is the subspecialty system that is difficult to manage and difficult to learn and influence by the individual officer. The system itself is composed of many administrative elements including a placement division, a division responsible for conducting a zero-based review of the entire subspecialty billet inventory, manpower claimants, sponsors, designator advisors, and assignment managers. Ninety-two percent of the survey sample viewed subspecialities as important for their careers. Yet, there is confusion among officers on the administrative steps they need to take to obtain a subspecialty; also, when to obtain one. For example, 65-percent indicated that it was important to obtain a subspecialty early in their careers. And yet, a Navy policy makes subspecialty experience obsolete after five years.

The ninth problem centers on dual career couples. Approximately 75-percent of the married General URLs are married to military men. These couples want to colocate. This desire provides: (1) career management problems for the Navy which must be concerned with mission requirements and (2) career development problems for General URLs who may want to advance in their careers rather than simply "be employed." The Navy currently has a policy that every reasonable attempt must be made to colocate couples.

The tenth problem is philosophical; i.e., there are disagreements about what direction the evolution of career management systems should take. For example, some community members want billets that are reserved exclusively for General URLs or want separate assignment managers. In this way, it is argued that women will be able to compete more effectively with their male counterparts. On the other hand, complaints are voiced that such approaches are an example of a "separate but equal" status that removes women from the Navy's mainstream and prevents them from competing successfully with male warfare specialists.

Policy Initiatives in the Career Area

In November 1984, the Navy promulgated a series of policies designed to alleviate some of the career management and development problems faced by the General URL community. One of the changes was that General URLs, rather than SWOs, would serve as assignment managers; i.e., General URLs would assign their own community. The exceptions to this policy are commanders who have screened for

command and CAPTS. These two sets of individuals will still be assigned by SWOs. The "new accessions" desk will also be manned by a SWO. In brief, the policies seem to contain elements of both philosophical positions mentioned earlier.

A second change was the institution of a two-career track, the Leadership/Subspecialty Track, which is the existing one, and a new Specialty Track. The former emphasizes both leadership and subspecialty billets, culminating in commanding officer positions. The new track emphasizes the opportunity to stay in a subspecialty track, eventually becoming a program manager rather than a commanding officer. However, a small percentage of those in the second track will become commanding officers of shore installations specializing in activities such as computer operations. Approximately one-third of General URL LCDRs will be accepted into the Specialty Track.

It is hoped that the two-career track better defines the billets that are needed to advance in the Navy, that it gives General URLs more options to fulfill their career goals, and that it provides the Navy with the specialized skills it needs to meet its requirements.

A third policy was aimed at stabilizing accessions into the community. While the numbers established operate within broad parameters, there is at least some structure and community control over this important issue. The previous community manager characterized this policy as the single most important change within the community.

A fourth policy was aimed at freeing up additional numbers of challenging and career enhancing billets for the General URL community. That is, 1,800 billets were identified that were reserved for warfare specialists, but which seemed within the capabilities of many General URLS. The goal, which was reached, was to be able to reclassify 300 of these billets so that General URLs would be eligible for them.

Before reclassification, General URLs were blocked, to a large extent, from obtaining billets that represented the operational opportunities needed for career advancement. The policy change addressed the complaint, previously quoted, regarding the poor quality of billets available to General URLs.

A fifth policy clarified the definition of leadership positions below the level of executive/commanding officer. That is, criteria were established for division officer and department head billets, and billets were appropriately recoded. This policy further defines the career path, thereby helping the individual officer to formulate and plot career strategy.

A sixth policy further defined the career path for officers just entering the Navy. The problem addressed by the policy was that new accessions were obtaining limited subspecialty and leadership experience in those situations where the Navy's needs were defined as being preeminent. The new policy stated that, under such circumstances, the first assignment will be split-toured or be a 2-year rather than a 3-year obligation.

A seventh policy was designed to ensure that General URLs, as they compete with warfare specialists, receive their fair share of leadership positions at the lieutenant-commander (LCDR) level. Although specific billets were not reserved, the policy stated that 75-percent of the LCDR executive officer and commanding officer shore billets will be reserved for the General URL community. The policy also dictated that the same type of arrangement be implemented as the community matures and sufficient numbers of commanders are available.

Finally, the recommendation was made to increase the fields in which General URL officers are assigned so that they have the breadth of experience to fill a wider range of leadership, HQ (i.e., headquarter) and subspecialty billets at the 06 (i.e., CAPT) level. A lead and an assistant agency were designated to determine how best to implement this recommendation.

Questionnaires will be mailed shortly and interviews conducted to determine the community's reactions to the actual and recommended changes discussed in this paper.

Attitudes, Preferences and Career Intentions of ROTC and Non-ROTC Students

Melvin J. Kimmel
US Army Research Institute for the Behavioral and Social Sciences[1]

The Reserve Officer Training Corps (ROTC), the Army's main source of officer personnel, finds itself in a dilemma, for it is being tasked to recruit and retain record numbers of officer candidates in a shrinking and more competitive market (Hertzbach et al 1985). The White male college-bound population has been its main source of cadets. However, changing market conditions will necessitate going beyond this traditional pool to include women and the growing number of adolescent Blacks and Hispanics (McNeil, 1983). To accomplish its expanded mission, ROTC must have information on the backgrounds and attitudes of these people in order to develop effective recruiting and training programs.

Over the past fifteen years, the ROTC Advertising and Media Division has relied upon comparative studies of ROTC cadet and noncadet college students to develop its programs (e.g., Armstrong, Farrell and Card, 1979; Card et al, 1975; Hicks et al, 1979; and Montgomery et al, 1974). These research efforts have generally drawn similar conclusions: (1) the influence sources that ROTC cadets and noncadets use to make career decisions are similar; (2) the ROTC and military-related attitudes of noncadets have become more positive since the Vietnam era; and (3) cadets and noncadets differ markedly on values, ROTC and military-related attitudes, preferences, and intentions, although some of these differences are true for only certain ethnic and gender subgroups.

The present effort continues this line of research with a more recent sample of ROTC and non-ROTC students. Its focus is on various socialization variables and military-related attitudes that may impact on one's decision to join ROTC and pursue a military career.

METHOD

Subjects. Usable data were gathered from 898 college students from 11 campuses with ROTC programs. The sample was composed of 427 first and second year ROTC cadets and 471 noncadet students. The frequencies for the ethnic and sex subgroups in ROTC and non-ROTC are presented below:

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Cadet	211	71	51	39	30	15
Noncadet	130	123	60	63	60	35

The majority of the sample (54%) were enrolled in southern colleges; 29% came from schools in the northeastern and mid-Atlantic regions of the country; 11% were enrolled in midwestern colleges; and 6% were from western schools. These percentages accurately reflect the geographic distribution of males and females in our sample, but they are not as characteristic of the cadet-noncadet breakdown or the geographic distribution of the different ethnic groups. The greatest discrepancies are found in the eastern and

---

[1]The views expressed in this paper are those of the author and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

southern college subsamples. Sixty-eight percent of the non-ROTC participants were from southern schools as compared to only 39% of the ROTC sample, while 18% of the non-ROTC sample and 41% of the cadets were enrolled in eastern colleges. An overwhelming majority of the Hispanics (86%) and Blacks (85%) in our sample came from southern colleges, while the White sample was more equally divided, with 34% from southern colleges and 42% from eastern schools.

Instrument and procedure. University staff members administered a slightly modified version of the 232-item "Career Attitude Survey" (Armstrong et al, 1979) during regularly scheduled class periods. The surveys were administered to ROTC cadets in MSI and MSII classes (The ROTC Basic Course) and to non-ROTC students enrolled in lower level required courses (e.g., English 101). The survey was composed of items on background characteristics, media preferences, education and career-related variables, and ROTC/Army knowledge and attitudes. The questionnaires took approximately 45 minutes to complete. All answer sheets were returned to a central location for coding, keypunching, 100% verification, and analysis. A 2 (Cadet-Noncadet) x 2 (Male-Female) x 3 (White-Black-Hispanic) factorial design was used to analyze main effects and interactions. The F-statistic was used for items associated with rating scales; and the x^2 statistic to analyze categorical data.

RESULTS

Career and Education Influences. A number of variables may influence one's decision to pursue a military career. Among these are the military attitudes and experiences of family and friends. When asked to rate on 5-point scales how favorably their parents and friends perceived the status of an Army officer career, the mean ratings of ROTC cadets were significantly higher than the non-ROTC student ratings on both perceived parental attitudes ($\bar{x}$=3.92 vs $\bar{x}$=3.41, F=35.76, p <.001) and the perceived attitude of their friends ($\bar{x}$=3.30 vs $\bar{x}$=3.06, F=13.92, p <01). Sex and ethnic differences were found in respondents' perception of their friends' attitude, but not with respect to their parents. Cadets and noncadet females perceived their friends' attitude as more positive than did the males ($\bar{x}$=3.42 vs $\bar{x}$=3.10, F=7.63, p <.01), and the Blacks' ratings ($\bar{x}$=3.88)were significantly higher than Whites ($\bar{x}$=3.18) or Hispanics ($\bar{x}$=3.23, F=10.94, p <.001).

When asked whether or not their parents, siblings, and friends had been in ROTC or the military, a significantly higher percentage of cadets than noncadets reported friends with ROTC experience (57% vs 50%, x^2=5.50, p <.05) and parents with military experience (69% vs 50%, x^2=19.25, p <.01). Ethnic differences also were found on ROTC and military experience variables. Fewer Hispanics (10%) than Whites (20%) or Blacks (14%) reported parents with ROTC experience (x^2=6.73, p <.05), while a greater percentage of Blacks reported siblings with ROTC experience (23% Blacks vs 13% each for Whites and Hispanics, x^2=11.97, p <.01). Blacks also had the highest percentages reporting military experience for siblings (28% Blacks vs 19% Whites and 15% Hispanics, x^2=11.97, p <.01) and friends (78% vs 69% and 68%, respectively, x^2=9.08, p <.05). Whites reported the highest percentage of parents with military experience (71% Whites vs 44% Hispanics and 38% Blacks, x^2=72.58 p <.001). The only significant sex difference that held for both cadets and noncadets was for friends with military experience, where the percentage of females was higher than males (73% vs 63%, x^2=3.86, p <.05). A higher percentage of

males than females reported parents with military experience, but the difference was more marked for cadets (72% vs 48%) than noncadets (58% vs 52% $x^2=6.79$, $p < .01$).

The students were also asked directly about the sources that influenced their decision on whether or not to participate in college ROTC. ROTC influence sources were measured by asking participants to indicate which of 14 potential sources influenced their decision. The resulting percentages for each source is presented in Table 1. Cadets and noncadets agreed on four of the most frequently mentioned influencers: Family, friends, personal beliefs, and career goals. Media advertising and ROTC unit requirements were among the least influential of the sources for both groups. As might be expected, a larger percentage of cadets than noncadets were influenced by ROTC instructors and other military personnel. The noncadets more often mentioned personal beliefs, career goals, and ROTC obligated service. Some significant ethnic and sex differences were found that characterized both cadet and noncadet groups. Whites chose career goals more often than Blacks or Hispanics. Blacks did not base their decision to join ROTC on personal beliefs as much as Whites or Hispanics, but were more influenced by ROTC recruiters and media advertising than the other ethnic groups. In addition, a larger percentage of males than females mentioned economic conditions as a factor in their decision.

Other observed sex and ethnic differences in ROTC influence sources were significant for cadets, but not noncadets. Specifically, a greater percentage of ROTC females than males said they were influenced by friends, (51% vs 45%, $x^2=11.84$, $p < .01$), teachers (18% vs 9%, $x^2=4.15$, $p < .05$), and ROTC instructors (48% vs 25%, $x^2=7.94$, $p < .05$), while more ROTC males than females were influenced by military lifestyle (24% vs 13%, $x^2=8.79$, $p < .01$) and personal beliefs (34% vs 25%, $x^2=4.15$, $p < .05$). The only race discrepancy between cadets and noncadets occurred in percentages reporting educational goals as influencers ($x^2=10.66$, $p < .01$). Within the ROTC group, a greater percentage of Whites (30%) than Blacks (16%) or Hispanics (11%) indicated that this influenced their decision. Within the noncadet group, on the other hand, the ethnic groups responded similarly.

ROTC and Military Attitudes. When asked how they felt about serving in the military, cadets and noncadets responded very differently ($x^2=122.02$, $p < .001$). The percentages of cadets and noncadets stating that they would serve if needed were about the same (52% for cadets and 46% for noncadets). However, a significantly greater percentage of cadets than noncadets stated that they felt a duty to serve (29% vs 6%) , while a much higher percentage of noncadets (48%) than cadets (19%) indicated that they had not given much thought to military service. In general, females and Blacks in both groups were less committed to military service. Forty-nine percent of the females in our sample indicated that they had not give much thought to military service; 41% said they would serve if needed; and only 10% believed it was their duty to serve. In contrast, only 25% of the males indicated giving no thought to military service, as compared to 54% that would serve if needed and 21% who believed it was their duty to serve ($x^2=59.10$, $p < .001$). With respect to ethnic group differences ($x^2=19.19$, $p < .001$), only 11% of the Blacks saw military service as a duty as compared to 19% of the Whites and 18% of the Hispanics, while 47% of the Blacks as compared to only 30% of the Whites and 30% of the Hispanics said they had not given much thought to military service. Fewer Blacks than Whites or Hispanics also indicated a willingness to serve if needed (42% vs 30% and 31%, respectively).

The more positive attitudes of cadets was also reflected when participants were asked to rate the attractiveness of ten aspects of an ROTC program on five point scales (see Table 2). Although cadets rated all ten aspects significantly higher than noncadets, the rank orderings of the program elements were relatively similar. Among the program elements rated most attractive were a guaranteed job after college, the scholarship program, program quality, and program activities (e.g., course modules, social functions, etc.). Obligated duty requirements, ROTC cadets, and program requirements were seen as the least attractive aspect of the program by both cadets and noncadets. The only exception to this consistent pattern was in the ranking of ROTC instructors, which was ranked first by cadets, but only fifth by noncadets. In general, Blacks and Hispanics rated these elements more positively than did the White subgroups. This pattern was significant for four of the ten program elements: Program image, program environment, ROTC cadets, and obligated duty requirements. Two of the sex differences were significant (activities and ROTC instructors), with females rating these elements more positive than the males. These ethnic and sex patterns characterized both cadet and noncadet groups in all cases except for the female Hispanic's attitude toward ROTC instructors. Cadet female Hispanics rated their instructors significantly lower than the other cadets, whereas in the noncadet group female Hispanics rated ROTC instructors more positively.

Similar patterns of results emerged when respondents were asked to rate 16 aspects of Army life (see Table 2). As with ROTC attitudes, cadet ratings were consistently higher than the ratings of noncadets, but their rank orderings were similar. Job security, officer responsibilities, and officer pay and fringe benefits were rated among the most positive by both groups, while personal freedom, prejudice, and Army living conditions were among the most unattractive elements of Army life. The largest discrepancies between the cadet and noncadet rankings were on "required mobility and travel," which was ranked fourth by the noncadets but only eleventh by cadets, and "required discipline", ranked seventh by cadets and twelfth by noncadets. Blacks in both groups rated all aspects of Army life more positively than the other ethnic groups, and Whites differed significantly from Hispanics only with respect to "job security," which they rated higher, and "personal freedom," which they rated lower. The only significant sex differences that characterized both cadet and noncadet groups were in attractiveness ratings of required travel and personal freedom, which were rated higher by females than males, and in their feelings about Army training, which was seen as more attractive by males than females. While ethnic and sex patterns were generally the same in cadet and noncadet groups, there was one exception for the element, "officer responsibility" ($F=4.87$, $p < .05$). Specifically, whereas the mean rating for ROTC males was higher than for ROTC females on officer responsibility ($\bar{x}=3.94$ vs $\bar{x}=3.83$, respectively), the reverse was true in the noncadet group, where the male means were lower ($\bar{x}=3.26$ vs $\bar{x}=3.36$).

DISCUSSION AND CONCLUSION

Results are remarkable similar to those reported in the earlier research efforts that compared cadet and noncadet characteristics. In comparison to noncadets, cadets as well as their family and friends continue to hold more positive attitudes toward ROTC and military service. Also consistent with the more recent efforts is our finding that noncadets are not as resistant to serving in the military as they had been during the Vietnam era. In fact,

there appears to be even less resistance now than at the time of the Hicks et al (1979) and Armstrong et al (1979) surveys. In addition, the present research parallels the earlier finding that cadets and noncadets rely on similar influence sources to decide whether or not to participate in an ROTC program, although cadets are still more influenced by military personnel, and noncadets by personal beliefs.

While the observed gender differences also are relatively similar to the Armstrong et al (1979) findings, the consistencies are less clear regarding ethnic comparisons. Both studies report similar ethnic group differences with respect to military service commitment, personal feelings about ROTC and the military, parental attitudes toward military service, and the military backgrounds of parents and friends. However, the two research efforts differ somewhat in ethnic group results for friends' attitude toward military service, parental ROTC experience, and the influence sources that students use to decide whether or not to participate in college ROTC. These discrepancies may indicate that the background experiences of college students have changed (at least, with respect to ethnic groups), or they may simply be the result of sampling error. Further research is needed to clarify this issue.

What are the implication of these findings to ROTC recruiting? First, the fact that noncadets are less resistant to entering military service than they were in the 1970's suggests that they may be more amenable to ROTC recruiting campaigns. Second, the results suggest that media advertising directed toward the potential recruit is not very effective. A better strategy might be to direct advertising programs toward what were found to be the major influence sources: Parents and friends. Finally, the large number of ethnic and sex differences observed in attitudes and potential influencers suggest that a multidimensional approach to ROTC recruiting is needed. Recruiting programs that are geared only to White males must be modified and expanded if ROTC is to attract the women and minorities that will make up a significant portion of the target pool in the 1990's.

References

Armstrong, T.R., Farrell, W. S. & Card, J. J. (1979). Subgroup differences in military-related perceptions and attitudes: Implication for ROTC recruitment. Research Report 1214. Alexandria, VA; US Army Research Institute for the Behavioral and Social Sciences.

Card, J. J., Goodstadt, B. E., Gross, D. E. & Shanner, W. M. (1975). Development of an ROTC/Army career commitment model. Washington, DC: American Institutes for Research.

Hertzbach, A., Gade, P. A. & Johnson, R. M. (1985). The 1984 ARI cadet decision-making survey: An overview of results. Working Paper 85-6. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Hicks, J. M., Collins, T. & Weldon, J. I. (1979). Youth aspirations and perceptions of ROTC/military: A comparison. Alexandria, VA; US Army Research Institute for the Behavioral and Social Sciences.

McNeil, I. (1983). Demographic Imperatives: Implications for Educational Policy. Washington, D.C.: American Council on Education.

Montgomery, J. R., McLaughlin, G. W., Pedigo, B. A., Mahan, B. T., and Associates (1974). Outlook on ROTC among high school, college, and ROTC students. Blacksburg, VA: Virginia Polytechnic Institute and State University.

Table 1. Percentage of Respondents Indicating Sources that Influenced Their Decision To Join or Not Join ROTC According to ROTC Membership, Ethnic Background and Gender

ROTC Influence Sources	ROTC Membership		Ethnic Background			Gender	
	Cadet	Noncadet	White	Black	Hispanic	Male	Female
Family	39	34	38	34	33	36	36
Friends	40	39	37	44	41	39	41
Personal Beliefs	32	45***	40	27	38**	37	36
Career Goals	29	38**	36	27	32*	34	32
Educ Goals	25	28	28	24	27	26	28
Military Lifestyle	20	30**	26	22	26	24	25
ROTC Instructors	32	5***	17	22	20	17	21
ROTC Recruiters	19	15	11	34	11***	16	18**
Econ Conditions	18	13	16	11	21	18	11*
Teachers/Counselors	12	12	11	15	10	11	15*
Military Personnel	16	7***	12	9	12	12	10
Obligated Service	3	11***	8	5	8	7	8
Media Ads	5	8	5	11	8**	6	8
ROTC Unit Requirements	3	6	5	3	3	5	4

* p<.05
** p<.01
*** p<.001

Table 2. Attractiveness of ROTC Program and Military Lifestyle According to ROTC Membership, Ethnic Background and Gender

ROTC PROGRAM Elements	ROTC Membership		Ethnic Background			Gender	
	Cadets	Noncadets	White	Black	Hispanic	Male	Female
Guaranteed Job	4.15	3.48***	3.76	3.79	3.80	3.79	3.80
Scholarship Program	4.08	3.45***	3.76	3.77	3.76	3.76	3.73**
Instructors	4.23	3.09***	3.65	3.63	3.62	3.73	3.49**
Quality	3.98	3.27***	3.60	3.66*	3.58	3.65	3.54***
Activities	4.02	3.11***	3.52	3.60	3.58**	3.66	3.36***
Environment	3.84	3.02***	3.32	3.61	3.48**	3.44	3.37
Image	3.67	3.01***	3.23	3.53	3.40**	3.33	3.32
Requirements	3.76	2.91***	3.28	3.44	3.32	3.35	3.26
ROTC Cadets	3.60	2.96***	3.17	3.46	3.38**	3.76	3.73
Obligated Army Duty	3.38	2.72***	2.95	3.26	3.04**	3.07	2.98
ARMY LIFESTYLE Elements							
Job Security	4.28	3.71***	3.99	4.17	3.71***	4.01	3.94
Responsibilities	3.91	3.31***	3.56	3.71	3.58	3.64	3.54
Pay/Fringe Benefits	3.92	3.26***	3.52	3.85	3.40***	3.52	3.67
Officer Quality	3.79	3.14***	3.39	3.67	3.37**	3.44	3.47
Army Goals	3.77	3.13***	3.34	3.71	3.40***	3.41	3.49
Recreation	3.64	3.11***	3.35	3.55	3.21*	3.38	3.36
Required Travel	3.50	3.15***	3.20	3.71	3.20***	3.19	3.52***
Relevance to Society	3.60	3.04***	3.31	3.46	3.09**	3.34	3.26
Discipline	3.63	2.86***	3.12	3.49	3.26*	3.28	3.14
Daily Activities	3.56	2.91***	3.11	3.53	3.22***	3.26	3.16
Training	3.60	2.85***	3.15	3.41	3.15*	3.34	3.01***
Personal Relations	3.38	2.92***	3.11	3.30	3.00*	3.14	3.14
Public Image	3.24	2.84***	2.91	3.42	2.95***	2.98	3.12
Living Arrangements	3.01	2.36***	2.49	3.06	2.82***	2.71	2.63
Prejudice	2.67	2.55***	2.59	2.79	2.44*	2.58	2.61
Personal Freedom	2.73	2.44***	2.43	2.90	2.70***	2.52	2.69*

* p<.05
** p<.01
*** p<.001

CAN MAINTENANCE DATA BE USED TO DEFINE TRAINING REQUIREMENTS?

*Michael Wagner
Dynamics Research Corporation
and
Major Martin Costellic
Air Force Human Resources Laboratory

In the 1970's the Air Force shifted the emphasis of maintenance training from the schoolhouse to the workplace. The average length of formal technical training decreased from 48 weeks to 25 weeks, while there was a corresponding increase in the OJT burden.

This new approach to training was expected to accomplish several things. First, it would get maintenance personnel out into the field more quickly. Second, it would allow recruits to spend more time training on equipment on which they would ultimately work. Finally, the emphasis on OJT would tie training more closely to the actual maintenance tasks that are performed in the field.

When some of the responsibility for technical training was transferred from formal schools to maintenance units, supervisors were given few resources to accomplish the additional training. As a result, the supervisor had little time to conduct OJT. Clearly, supporting mechanisms are necessary to help OJT supervisors achieve their expanded training mission.

One mechanism designed to support the unit training needs is the Field Training Detachment (FTD). Operated by ATC, FTDs smooth the transition from formal training to the unit. After completing generalized technical courses, airmen are assigned to a base whereupon they are sent to an FTD. There they receive equipment - specific training on the actual equipment maintained at the base.

In order to determine training requirements for OJT (e.g., FTD) data is required that provides accurate and timely descriptions of the specific tasks which comprise a job (including equipment references), how time is allocated among these tasks, and how well these tasks are performed.

The purpose of the Training/Job Requirements System contract with AFHRL/ID was to assess the feasibility of using automated maintenance data collection (MDC) systems to develop <u>accurate</u>, <u>objective</u> and <u>real-time</u> job descriptions of maintenance jobs.

Maintenance-Based Job Description Reports

Data on all maintenance activities performed on each aircraft are collected as part of the MDC system. With few modifications, this system can also keep track of personnel who perform the maintenance actions and could be used to develop historical summaries or descriptions of the maintenance actions that constitute a job. The principal objections to using data from MDC in the past to define jobs have been that the data are (1) incomplete and (2) inaccurate. However, with the advent of automated maintenance data collection systems such as the Centralized Data System (CDS) on the F-16 aircraft and the Automated Maintenance System (AMS) on C-5A aircraft, these objections are no longer valid. Specifically:

(1) Automated maintenance data collection systems ensure that data is complete and accurate. Work order generation and tracking ensures that maintenance actions are recorded; on-line editing produces accurate and complete entry of maintenance information; a centralized system contains daily data from all maintenance sites;

(2) Maintenance data consists of a set of standard task elements which refer to equipment involved (work unit codes) and actions taken (action taken codes). In addition, the clock time for specific maintenance actions, although influenced by the general workload level, provides accurate estimates of relative task times;

(3) Maintenance task data can be collected which identifies each crew member involved in a maintenance action. This data can be obtained without imposing an additional burden on maintenance personnel.

(4) Maintenance data on individuals can be aggregated over time to construct reports summarizing all maintenance actions performed by an individual. Reports can be constructed as a function of skill level, AMU, or base.

(5) The computer data base and terminals for collecting these data are already in place, at least for the F-16 aircraft.

A small portion of a maintenance job description report that can be generated using maintenance data is shown in Figure 1.

AFSC: 328x8
Base: MacDill AFB, 56th TTW
Time Period: 85014 to 85054

			SUM TOTAL TIME SPENT	FREQ.	MEAN CLOCK TIME	SUM PERCENT TOTAL TIME
			2352.23	1844.00	1.28	100.00
WUC	**AT**	**CDS DESCRIPTION**	**TOTAL TIME SPENT**	**FREQ.**	**MEAN CLOCK TIME**	**PERCENT TOTAL TIME**
74A00	L	Adjust fire control radar set.	42.68	33.0	1.29	1.81
74A00	Q	Installed fire control radar set.	4.00	2.0	2.00	.17
74A00	R	R&R fire control radar set.	6.60	5.0	1.32	.28
74A00	V	Clean fire control radar set.	8.00	4.0	2.00	.34
74A00	X	T/I/S fire control radar set.	61.54	63.0	.98	2.62
74A00	Y	Troubleshoot fire control radar set.	181.81	100.0	1.82	7.72

Figure 1. F-16 CDS Job Description Report Excerpt

This maintenance data-based report provides a listing of specific tasks performed and how time is allocated among these tasks.

A second type of report, shown in Figure 2, provides performance data. This report, called a WUC Removal Action Performance Report, identifies how often avionics LRUs removed at the flightline are later found to be serviceable when they are tested in the Avionics Intermediate Shop (AIS). This information can be generated for either a base or a work center and will identify which serviceable LRUs are frequently removed unnecessarily. This information can be used to decide if additional resources need to be allocated to the training of troubleshooting particular equipments.

AFSC 32X3X

Time Period 8501 to 8503

Base/WC MacDill/AC

(1) WUC	(2) Number Received In Shop	(3) Number Processed	(4) CND S(%) BCS	(5) S(%) Repaired	(6) S(%) NRTS
Z11BJ	3	4	0(0)	3(75)	1(25)
51BA0	8	8	0(0)	4(50)	4(50)
51BB0	0	0	0(0)	0(0)	0(0)
51EB0	0	0	0(0)	0(0)	0(0)
51FA0	8	8	3(37.5)	3(37.5)	2(25)
74AA0	7	7	0(0)	1(14)	6(86)
74AB0	26	26	11(42)	12(46)	3(12)
74AC0	1	1	0(0)	3(100)	0(0)
74AD0	6	6	2(33)	3(50)	1(17)
74AF0	7	5	1(20)	3(60)	1(20)
74BA0	17	8	0(0)	7(87.5)	1(12.5)
74BC0	5	5	0(0)	5(100)	0(0)
74CA0	3	3	0(0)	3(100)	0(0)
74DA0	14	14	3(21)	11(79)	0(0)
74DD0	11	11	0(0)	2(18)	9(82)
74EA0	9	6	0(0)	6(100)	0(0)
74EB0	0	3	0(0)	2(67)	1(33)

Figure 2. Sample WUC Removal Action Performance Report

Users of these reports could include on-the-job training (OJT) supervisors; base training analysts; maintenance training managers at TAC; and analysts at ATC, MPC, and OMC involved in describing maintenance jobs, developing training standards, and evaluating training effectiveness.

Linkages to Current Job Descriptions and Training Standards

Job descriptive information derived from maintenance data support the definition of training requirements only if this information can be linked to current job descriptions and training standards. Only after such linkages are created can the existing job descriptions and training standards be evaluated and subsequently revised. Thus, in order to realize the full benefit of collecting maintenance data on individuals, linkages must be established for all the AFSs on which maintenance data is collected. Figure 3 provides an illustrative example of how linkages are established among maintenance tasks (F-16 CDS), occupational survey tasks (OS), and specialty training standards (STS).

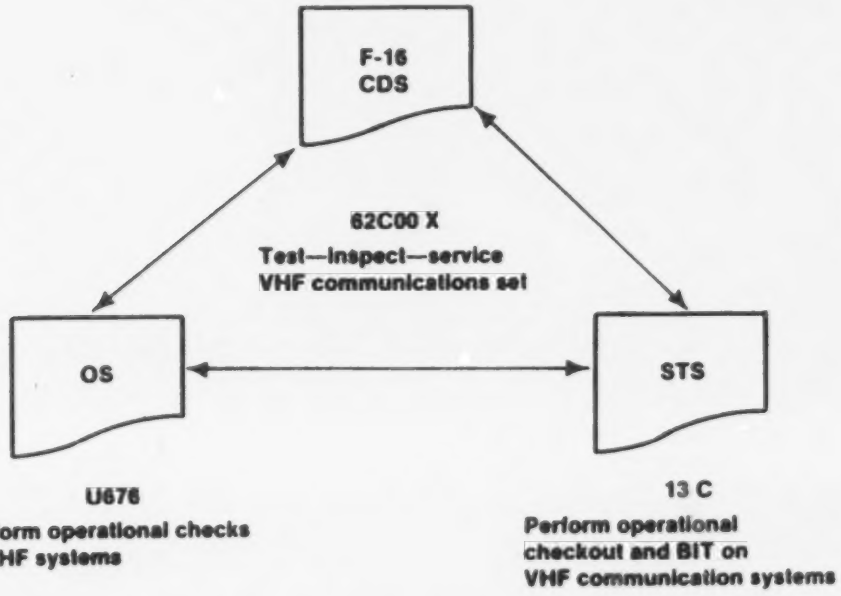


Figure 3. Task Linkages

Summary

Several capabilities were demonstrated in this study which have wide applicability. These capabilities include using automated maintenance data collection (MDC) systems to collect task data on maintenance personnel, linking these maintenance tasks to current AFS job descriptions (i.e., occupational survey reports) and specialty training standards, and developing reports which can be used to construct training requirements tailored to specific groups (i.e., bases) and to identify training deficiencies.

TRAINING EMPHASIS: THE BEST TASK FACTOR AVAILABLE

Lieutenant David L. Hardy, Lieutenant Colonel Charles D. Gorman,
and Dr. Walter E. Driskill
USAF Occupational Measurement Center

PURPOSE

This paper illustrates three important facets of the training emphasis (TE) factor. First, TE ratings provide the best data for establishing training priorities in comparison to the other task factors frequently cited as needed for curriculum decisions. Second, the reason training emphasis is the best task factor for establishing training priorities is the significant correlations that exist between the TE ratings and other task factors. Third, TE data are sufficient for instructional systems development (ISD) decisions. Following a brief historical summary of the training emphasis factor and a discussion of the just mentioned facets of TE data, a brief description of how TE technology is applied by the USAF Occupational Measurement Center (OMC) will be provided.

HISTORY AND BACKGROUND

The training emphasis task factor developed from a long, detailed, and complex research program. Christal (1970) first proposed the gathering of task factor information using subject-matter experts as the pool of raters from which samples could be drawn. Mial and Christal (1974) and Mead (1975) conducted the initial research using the policy-capturing approach to successfully predict training priorities. In addition, Mead's research was particularly hopeful in suggesting an intimate integration of ISD practices, occupational survey data, and curriculum design. In 1977, Stacy, Thompson, and Thomson reported that standard occupational survey techniques were reliable for the collection of task training factors.

The first major research specifically on training emphasis was reported by Ruck, Thompson, and Thomson (1978). This report established the ground work for use and analysis of TE data. Some of their recommendations were: that TE data should be collected and not predicted; that training emphasis ratings be collected while the routine collection of task delay tolerance and consequences of inadquate performance task factors should be discontinued; and that ratings be separately collected for each Air Force specialty.

The TE task factor research was released for operational use to the USAF Occupational Analysis Program, USAFOMC, Randolph AFB, Texas, in 1979. Driskill and Mitchell (1980) mentioned the exstensive gathering of TE data for use by technical training curriculum developers and training managers. In 1981, Staley and Weissmuller presented a paper on interrater reliability in the CODAP programs. They suggested training emphasis is a stable task factor in noncomplex specialties, but that more research needs to be done on the application of TE data to complex specialties. Jansen (1982

and further in 1985) tackled the issue of complex specialties and proposed a common rating policy (CRP) approach which is adequate for all but a few specialties.

TE vs OTHER TASK FACTORS

Before suggesting an integration of the TE task factor into a prominent position in ISD theory and practice, a discussion of the correlations between TE and a number of other task factors investigated by Goldman (1985), and Ruck, Thompson, and Stacy (in preparation) is necessary.

Goldman, in a different approach, both statistically and methodologically, showed the same conclusions alluded to by Ruck, et al (1978), that the TE task factor is the best task factor available and that other "training" factors are highly correlated with TE. The factors used in Goldman's research are identified in the Instructional Systems Development (ISD) Eight Factor Model. In addition to the TE factor, those factors included:

1. Percent of members performing
2. Average percent of time spent by members performing
3. Task Learning Difficulty (LD)
4. Consequences of Inadequate Performance (COIP)
5. Task Delay Tolerance (TDT)
6. Probability of Deficient Performance (PDP)
7. Immediacy of Performance (IP)
8. Relative Frequency (RF)

The conclusions reached by Goldman suggest that instead of nine separate factors, there is only one clearly defined training factor; and, in terms of predicting critical vs noncritical tasks, rather than nine factors, there is really only TE. Also, by collecting a minimum number of task factors, efficiency in data gathering and analysis is significantly enhanced. These conclusions are a result of the high correlations found between the TE task factor and the other factors as shown in Table 1.

Research conducted by Ruck, Thompson, and Stacy (in preparation), has direct impact on the use of task factors in the ISD program. Jansen (1985) states, "the utility, reliability, and validity of training emphasis ratings in terms of ISD theory have been demonstrated", by the aforementioned authors. The thrust of their research was the development of a task training emphasis scale and some training priority equations. The following task factors were used in their study:

1. Training Emphasis
2. Probable consequences of inadequate performance
3. Task delay tolerance
4. Learning difficulty
5. Percent members performing
6. Percent time spent
7. Task grade-level index

The results of Ruck, Thompson, and Stacy's analysis shows a very positive correlation between TE ratings and several other task factors (see Table 2). The conclusions reached by Ruck, et al, in relation to other task factors used are: that TE ratings are "construct valid" because they can be predicted by using ISD training factors; TE ratings are reliable since supervisors give their ratings independently and have high agreement with one another; and, the process used in their study for arranging task lists in order of training priority should be adopted by the training community to improve training. Further, they recommend that TE ratings be routinely collected and "that the consequences of inadequate performance and task delay tolerance factors be collected for those specialties for which they are of special interest, since recommended TE ratings include consideration of these factors."

TE AND ISD

The ISD program for the Air Force is operationally defined in AFP 50-58 (1978), then outlined in AFM 50-2 (1979). From all ISD sources, there are seven stated task factors; they are:

1. Percent members performing
2. Number of members performing
3. Consequences of inadequate performance
4. Task delay tolerance
5. Task learning difficulty
6. Frequency of performance
7. Training development time

From previous discussion in this paper, it has been established that training emphasis is the best task factor available, and other factors are, perhaps, redundant--mainly because TE ratings include consideration of most of the other task factor variables. The key point of Ruck, et al's, report is that training emphasis should be the most extensively used task factor when using the ISD approach to make training decisions. Indeed, their research provided several CODAP computer programs that allow for presentation of TE data in very effective, usable formats.

CURRENT APPLICATION OF TE TECHNOLOGY BY USAFOMC

Since the release of TE technology for operational use, USAFOMC routinely has collected TE ratings as a standard part of conducting occupational analysis. After the ratings have been entered into the computer, the CODAP program REXALL is used to identify and eliminate deviant raters. This process is repeated until an acceptable interrater reliability for a single rater is at least .20, and the interrator reliability for all raters is .90 or more. After achieving these minimum levels of acceptability, any remaining deviant raters are individually reviewed for retention or rejection, based on the correlation of their ratings with mean ratings for the total group. At this point, raters that have been eliminated are examined to determine if there are any systematic similarities among them. Similarities may suggest the presence

TABLE 1

Correlations Between Training Emphasis and Eight Other Task Factors

MOS	FACTOR 1	2	3	4	5	6	7	8	9
11B	1.0	-.45	.81	.78	.86	.91	.77	.86	.76
63H	1.0	.68	.77	.84	.90	.76	.87	.62	.59
91C	1.0	-.27	.27	.82	.73	.78	.76	.83	.76
76Y	1.0	-.50	.25	.71	.85	.75	.77	.82	.73
95B	1.0	-.57	.77	.87	.91	.87	.85	.85	.70
12B	1.0	-.52	.71	.83	.78	.74	.64	.71	.64

NOTE 1:
1 = Training Emphasis
2 = Learning Difficulty
3 = Consequences of Inadequate Performance
4 = Task Delay Tolerance
5 = Probability of Deficient Performance
6 = Immediacy of Performance
7 = Relative Frequency
8 = Percent of Members Performing
9 = Average Percent of Time Spent by Members Performing

NOTE 2: Table adapted from Tables 1 and 2 in Goldman's (1985) Report.

TABLE 2

Intercorrelations Between Training Emphasis and Six Other Task Factors

DAFSC	FACTOR	1 2	3	4	5	6	7
293X0		38	-82	00	82	75	-66
304X0		44	-80	11	73	54	-71
304X4		20	-35	03	77	71	-21
328X3		39	-52	-49	77	71	14
423X1		63	-79	-11	78	70	-47
423X4		43	-27	-20	88	80	-25
427X5		89	-95	-49	88	77	-84
472X2		15	-12	-30	87	78	-43
531X5		72	-80	02	88	82	-78
552X5		31	-32	-35	79	65	-73
631X0		36	-75	-64	74	63	-61
651X0		41	-72	00	60	54	-65
672X1		43	-24	-17	70	63	-36
672X2		00	-71	-43	85	75	21
902X0		33	-59	-40	75	61	-67
906X0		28	-39	-04	48	43	-43
911X0		12	-76	-29	74	65	-06
981X0		28	-79	-61	85	79	-77

NOTE 1:
1 = Training Emphasis
2 = Probable Consequences of Inadequate Performance
3 = Task Delay Tolerance
4 = Learning Difficulty
5 = Percent Members Performing - First Job
6 = Percent Time Spent - First Job
7 = Task Grade-Level Index

NOTE 2: Table adapted from Appendix D in Ruck, et al, report (in preparation)

NOTE 3: Values are decimal values between 1.00 and -1.00

of multiple policies in the career ladder. Also, if more than 40 percent of the raters in the original sample are eliminated, a separate REXALL should be run, even though there are no apparent systematic similarities, to determine if another rating policy exists.

Once an acceptable REXALL has been achieved, the process of identifying deviant tasks begins. To do this, the mean of the task standard deviations (AVESD) and the standard deviation of the standard deviations (SDSD) (multiplied by 1.5) are added together. Tasks whose standard deviations exceed this value are defined as deviant tasks. These identified deviant tasks are then examined for systematic similarities. If the deviant tasks are random, then the data can be used with confidence. If the deviant tasks are systematically related, then the data are examined for multiple policies or the data will be used as they are with a caution to the user that within certain areas the raters disagreed as to the amount of emphasis that sould be placed on training.

SUMMARY

This paper briefly discussed the training emphasis task factor. To begin, background to the evolving technology of training emphasis was capsulized. This established the ligitimacy of the training emphasis ratings. After establishing some of the initial research results, a comparison of training emphasis to a number of other training-related task factors was made. This comparison showed TE ratings to be superior to the other task factors because of its predictive value and consistent incorporation of other task factors within its ratings. Next, a short discussion pointed out the importance of training emphasis ratings in the ISD training development process. Finally, a brief description was given on how the USAFOMC is currently applying TE technology. The purpose of this paper is to urge the incorporation and integration of the training emphasis rating as the single most useful task factor available to the training community, and to those who are using the Instructional Systems Development approach for making training decisions.

REFERENCES

Air Force Manual 50-2. Instructional System Development. Washington, D.C., Deprtment of the Air Force, 25 May 1979.

Air Force Pamphlet 50-58. Handbook for Designers of Instructional Systems (5 vols.). Washington, D.C. Department of the Air Force, 15 July 1978.

Christal, R. E., Implications of Air Force Occupational Research for Curriculum Design. In B. B. Smith and J. Moss, Jr. (Eds.), Report of a Seminar: Process and Techniques of Vocational Curriculum Development. Minneapolis, MN: University of Minnesota, 1970.

Driskill, W. E., and Mitchell, J. L. The USAF Occupational Analysis Program: An Evolving Technology. Proceedings of the 22nd Annual Conference of the Military Testing Association, Toronto, Canada, October 1980.

Goldman, L. A. Determining the Best Set of Training Factors. Proceedings of the Fifth International Occupational Analysts Workshop, Randolph AFB TX, May 1985.

Jansen, H. P. Identification of Rating Policies in Training Emphasis Task Factor Data. Proceedings of the 24th Annual Conference of the Military Testing Association, San Antonio, TX, November 1982.

Jansen, H. P. Training Emphasis Task Factor Data: Methods of Analysis. (AFHRL-TR-84-50). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, May 1985.

Mead, D. F. Determining Training Priorities for Job Tasks. Proceedings of the 17th Annual Conference of the Military Testing Association, Indianapolis, IN, September 1975.

Mial, R. P., and Christal, R. E. The Determination of Training Priority for Vocational Tasks. Proceedings, Psychology in the Air Force Symposium. USAF Academy, April 1974.

Ruck, H. W.; Thompson, N.A.; and Stacy, W. J. (In preparation). Development of a Task Training Emphasis Scale and Training Priority Evaluations. Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Ruck, H. W.; Thompson, N. A.: and Thomson, D. C. The collection and Prediction of Training Emphasis Ratings for Curriculum Development. Proceedings of the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK, October-November 1978.

Staley, M. R., and Weissmuller, J. J. Interrater Reliability: The Development of an Automated Analysis Tool. Proceedings of the 23rd Annual Conference of the Military Testing Association, Washington, D.C., October 1981.

TRAINING EFFECTIVENESS ANALYSIS OF DRAGON SIMULATORS

Mr. Walter G. Butler

US Army TRADOC Systems Analysis Activity

The requirement to provide combat units with soldiers who are qualified and proficient in the use of the DRAGON antiarmor weapon system currently entails the use of training ammunition, ranges, and troop support facilities. The high cost and limited availability of these resources are constraints on the effectiveness of the institutional training program to produce such soldiers. In addition, current training devices limit the provision of a broad range of realistic target engagement conditions.

This paper presents the results of an evaluation of the cost and training effectiveness of three alternative DRAGON training programs. The programs were almost identical in structure, differing only in the training device(s) used to support the program. Three training devices were involved; the launch effects trainer (LET), the launch environment simulator (LES), and the simulated tank antiarmor gunnery system - DRAGON (STAGS-D).

The LET and LES were already being used for DRAGON training; the STAGS-D was in development. These devices were combined in the following mixes to produce the alternative training programs:

- o LET and LES (base case)
- o STAGS-D alone
- o STAGS-D and LES

SYSTEM DESCRIPTIONS

DRAGON is a medium range wire-guided antitank weapon. It is fired by one man from the right shoulder with a bipod supporting the front end of the launcher. The two major components of the DRAGON are the round and the tracker. The round consists of a missile prepacked in its launcher. The tracker is mated to the round before firing. After firing, the empty launcher is discarded or destroyed and the tracker is used on the next round. The DRAGON can be fired at night by replacing the day tracker with the thermal night tracker (AN/TAS-5).

To engage a target with DRAGON, the gunner places the crosshairs in the tracker sight on the target and depresses both a thumb safety switch and a palm firing switch. There is a 0.6 second delay before the rocket motor fires to eject the missile from the launcher. The gunner continues to track the target until missile impact. During flight an infrared (IR) flare at the end of the missile allows the tracker to determine the relative position of the missile to the aim point. The tracker sends commands through the wire link to 30 pairs of thruster motors on the missile to correct deviations from the aim point. When corrections are needed, a single pair of thrusters will fire. These thrusters cannot be fired again.

The STAGS-D is designed to provide training under day or night conditions with single or multiple targets at up to three different ranges. Targets are scale-model combat vehicles mounted on carrier assemblies that can move on target tracks from left to right or right to left. Each target vehicle carries an infrared emitting diode so that an IR sensitive camera can determine its location. A target can be stationary or moving at a prescribed rate with intermittent stops if desired. The target vehicle can also rotate to present a head-on target and return simulated fire on the gunner position. Various levels of battlefield obscuration can be electronically inserted in the gunner sight to provide realistic firing conditions.

The LET consists of a facsimile of the DRAGON launcher, a performance monitoring set, and a towed target panel with an infrared source. A standard day or night tracker is attached to the LET launcher to complete the system. The LET is designed to teach firing and tracking procedures and to provide the student gunner with a representation of several effects he would experience while firing a live DRAGON. When conducting a LET engagement the gunner must depress the safety and firing switches and then wait through the 0.6 second delay as with the live missile. The launcher assembly then fires a standard NATO 7.62mm grenade cartridge at the rear of the launch tube. This provides launch noise and recoil simulation and activates a weight shift mechanism to simulate the weight loss of a missile leaving the launcher. The noise provided by the LET is not as loud as that of an actual firing, neither is there any heat or obscuration from smoke and debris. The gunner tracks the target panel at an actual range of about 250 meters. The LET simulates target ranges from 100 to 1000 meters (in 100-meter increments) by requiring the gunner to track for different elapsed times before the device terminates the engagement.

The LES is a recent addition to DRAGON training. The device operates under the same concept as the LET in that the gunner learns firing and tracking procedures and his performance is displayed on the monitoring set (same one used with the LET). The difference between LET and LES is that the LES provides much closer simulation of the noise, heat, and obscuration produced by a live DRAGON firing. The launcher has been modified to replace the grenade cartridge assembly with a chamber in which an explosive mixture of Methylacetylene Propadiene, Propane Propylene (MAPP) gas and oxygen is retained. A separate control box is used to house the MAPP gas and oxygen containers and the controls needed to fill the launcher chamber with the proper mix of gases. When the LES is fired a glow plug ignites the gas mixture to produce noise, recoil, heat, and smoke that approximates that of an actual DRAGON firing. Further obscuration is produced by fragmentation of fibrous endcaps on both ends of the launcher and by dust and debris raised by the explosion. The weight loss sensation is simulated by the downward deflection of some of the blast force at the rear of the launcher. Because of the noise level intensity, the Surgeon General of the Army has directed that no soldier will fire the LES more than five times per day or more than 20 times per year.

METHODOLOGY

The sample of students for the study consisted of 150 soldiers who participated in the DRAGON gunner course during one station unit training (OSUT) for MOS 11B at Fort Benning, GA. Approximately thirty soldiers were

taken from each of five classes conducted from 9 October through 15 November 1984 to complete the sample. The soldiers were randomly assigned to one of the three training programs under evaluation. Eighty-seven percent of the soldiers were active Army enlistees with the remainder being in the National Guard. No distinctions were made between these types of soldiers in the analysis since all had been in service for less than ten weeks and had received the same basic training.

The instruction group consisted of nine members of the Infantry Training Group (ITG) at Fort Benning. These soldiers were qualified DRAGON instructors who had been teaching in the DRAGON gunner course. They had received a one week training program on the operation of the STAGS-D before the test start.

The current DRAGON gunner course at Fort Benning was the base case. The course is a 40-hour program consisting of a combination of lecture and hands-on training, followed by two gunnery tables used to determine the qualification status of each student. Both the LET and LES are used for hands-on training. The qualification tables require the soldier to engage a series of targets under various conditions using the LET and the LES.

The remaining two alternative training programs were structured in the same way as the base case program with appropriate substitution of training devices. Alternative 2 replaced both the LET and LES with the STAGS-D, while alternative 3 included the LES and used the STAGS-D in place of the LET. The alternatives are referred to as the LET/LES POI, the STAGS pure POI, and the STAGS/LES POI.

The study required data concerning soldier proficiency with an actual DRAGON missile and with each simulator used during training and soldier perceptions of the alternative training programs.

The proficiency of the soldiers with the DRAGON was assessed in live firing exercises at the end of each week. The live firings were conducted on South Galloway Range at Fort Benning using inert missiles. The target was a manned evasive target tank (METT) with a driver and tank commander aboard. The vehicle travelled alternately right to left and left to right at 5 miles per hour approximately 830 meters from the firing point.

Each live missile firing was recorded on video tape by TRASANA personnel using the gunner aiming sensor (GAS) system. The GAS consists of a six-power objective optical lens, a six-foot fiber optic bundle, a color video camera, a video tape recorder, and a monitor. The objective lens, which produces the same magnification as the DRAGON tracker, is attached to a base plate bonded to the tracker. The fiber optic bundle carries the image from the lens through an 800 x 1000 array of 10-micron fibers to the camera enclosed in a shock-resistant case. The camera transmits the image to a recorder located up to 60 feet away. The image is stored on tape and simultaneously displayed on a monitor for realtime viewing of the engagement. The engagement can be replayed to assist in determining hit or miss and to distinguish between a missile malfunction and a gunner error in case of a miss.

The gunner aiming sensor provides a reticle in the image transmitted to the camera. This reticle can be boresighted to the tracker reticle by adjustment

mechanisms located on the bracket that attaches the lens to the base plate on the tracker. The alignment of the GAS and tracker reticles was checked before and after each shot and proved quite stable as only small corrections were necessary after several firings.

The presence of the reticle on the video tape of a live firing permitted a detailed analysis of gunner performance. The Data Sciences Division, National Range Operations Directorate at White Sands Missile Range, New Mexico, analyzed the tapes using manual and automated techniques to provide the magnitude of gunner aiming error at various points during the engagement, the times at which certain critical events occurred, and other data describing the engagment.

Data relating to soldier proficiency with the simulators were collected during the two qualification firing tables conducted at the end of each class. These data include the number of hits recorded on each of the qualification tables, the number of hits with each simulator, and the qualification status of each soldier as a result of his performance on the tables.

The perceptions of the soldiers concerning the simulators and training programs were obtained through surveys and written comments after the soldiers had fired the live DRAGON.

RESULTS

The video tapes of the live firings were analyzed to answer four questions about each soldier's performance during his DRAGON engagement. First, did the soldier achieve a successful launch by regaining control of the missile after undergoing the launch effects (heat, noise, shock, etc.)? Secondly, did the soldier keep the missile in the air long enough to reach the target; in this case, about eight seconds? Thirdly, did the soldier fly the missile through a rectangle containing, but slightly larger than, the target; that is, did he come "close" to a hit? Finally, did the soldier hit the target? A comparison of the results for the soldiers in each of the three training programs showed no statistical difference in performance among the groups. In fact, the largest difference between any two groups in any area was six percentage points. This analysis did reveal several points of interest concerning the engagements. Nearly half of all the target misses were caused by unsuccessful launches. Almost all soldiers who achieved a successful launch flew the missile "close" to the target. Many of the missiles that came close to the target but missed did so because of excessive tracker movement by the gunner during the last few seconds of missile flight. These observations provide the developers and providers of DRAGON gunner training with specific areas in which increased emphasis can produce significantly improved performance.

The video tapes also allowed examination of the maximum deflection due to launch effects of the tracker reticle from the ideal aim point on the target. Measurements were taken in both the horizontal and vertical planes. Again, no differences were observed between the groups. It was noted, however, that those soldiers who hit the target tended to have allowed a smaller deflection after launch in the horizontal plane than those soldiers who missed. This tendency did not hold in the vertical plane. DRAGON gunner training has long

emphasized control of the tracker in the vertical plane to reduce the possibility of grounding the missile during or just after launch. The results noted above suggest horizontal control is at least as important in ultimately achieving a target hit.

These analysis results show that no differences could be found between the training effectiveness of the three training programs. Similarly, the twenty year life cycle costs of the programs were found to be nearly the same with the largest cost difference being ten percent.

The results of the qualification exercises using the various simulators were compared to the live-fire results to determine if any relationships could be found between soldier performance with the devices and performance with the DRAGON in terms of target hit or miss. The twenty qualification shots are divided into two tables (called Table III and Table IV) of ten shots each. The number of shots with each simulator is shown below for each training program:

	LET/LES	STAGS Pure	STAGS/LES
Table III	10, LET	10, STAGS	10, STAGS
Table IV	5, LET 5, LES	10, STAGS	5, STAGS 5, LES

The qualification standard used in previous DRAGON gunner courses was at least eight hits on each firing table. When this standard was applied to the results of the qualification firings of the soldiers involved in the study, thirty-one (62%) of the soldiers in the LET/LES program were deemed qualified while no one in the STAGS pure program and only two (4%) of the STAGS/LES soldiers were qualified. The qualification standard was determined to be inappropriate for the latter two programs, since all the programs had been shown to be equally effective. Considering only the soldiers in the LET/LES program, the existing qualification standard correctly predicted the hit/miss outcome of the live-fire engagement for 60% of the soldiers. By contrast, an alternative standard that ignored Table III and required at least nine hits on Table IV correctly classified the live-fire results of 68% of these soldiers. Still another standard that ignored the LET altogether and required at least four hits out of the five LES engagements on Table IV provided 70% correct classification.

For soldiers in the STAGS pure program, no relationships were found between performance on the STAGS and live-fire performance.

For soldiers in the STAGS/LES POI, a relationship was found that involved a requirement of at least one hit with each simulator and a total of at least four hits among the twenty qualification shots. This requirement provided 68% correct classification of live-fire performance.

The most notable result of this area of the analysis is that the LES seems to provide some relationship between performance with the simulator and performance with the DRAGON itself.

The compilation and analysis of the survey responses and written comments revealed the following soldier perceptions:

- o all simulators, except possibly LES, needed more launch effects (noise, heat, smoke, etc.)
- o STAGS was much more sensitive than DRAGON to gunner movement
- o STAGS hurt confidence and morale
- o Difficulties with STAGS and DRAGON seem easier

These perceptions indicate strengths and weaknesses of the STAGS simulator. Apparently the over-training provided by the STAGS in teaching steady tracking techniques caused the soldiers to feel they had good control over the actual DRAGON. The increased sensitivity of the STAGS that provided the over-training, however, also caused the soldiers to achieve far fewer hits during practice and qualification than did the soldiers who trained with the LET/LES. The soldiers hit 78% of all targets engaged with the LET or LES and 30% of all targets engaged with STAGS. This led to confidence and morale problems among the soldiers training on the STAGS.

SUMMARY

The analysis of the three DRAGON training programs produced the following results:

- o the programs produced essentially the same levels of performance in the test soldiers
- o the programs cost essentially the same
- o performance with the LES provided information about performance with DRAGON

The following results concerning DRAGON gunner training in general were also observed:

- o half of the live-fire misses were due to unsuccessful launches
- o most of the remaining misses were due to tracking instability during the last few seconds of flight
- o tracking control in the horizontal plane during launch was just as important as vertical control

Career Involvement of Recent West Point Graduates

Jerome Adams
United States Military Academy
West Point, NY 10996-5000

John D. Richards
Academy of Health Sciences
Fort Sam Houston, TX 78234-6100

In today's rapidly changing society, it is important to stay abreast of current developments, especially within the context of significant military attitudes which could affect the readiness of our Forces. Much has been written in recent years about career decision-making, planning strategies, career involvement, dual careers (Adams 1980; Davenport 1984; Hall & Hall, 1979), and overall commitment and adjustment to career and organization. Moreover, Moskos (1977) has noted changes from an institutional to an occupational organizational orientation among members of the military. Some suggest that today's military has less of the rational organizational devotion and dependency, and more of the entrepreneurial protean man (Hall, 1976; 1979). Hall (1979) also distinguishes between moral and calculative organizational commitment, with a shift being analogous to the phenomenon observed by Moskos (1977). In terms of personal decisions the concern with career planning and strategies utilized in making those decisions also has changed.

The present paper reports recent research pertaining to career attitudes and proclivities of male and female West Point graduates (Adams 1984, Adams & Yoder 1984). The data to be reported were imbedded within the context of larger, world-wide study of early adjustment experiences of US Military Academy graduates. Specifically, strategies of career planning, degree of career involvement, and overall commitment and adjustment to the Army were analyzed.

METHOD

Respondents. The participants were selected based upon stratifying by sex, work specialty and geographic assignment. The samples reflected the same proportions of branch specialty and location as the entire population of graduates. A ratio of three men were selected for each female respondent. The two year interviews were conducted with 116 members of the class of '80 and 104 members of the class of '81 at locations in CONUS, Korea, and Germany. A breakdown by gender and location is presented in Table 1.

Procedure. The two year protocol was the first to include extensive questions on social aspects of Army, Army life, and the interaction of an Army career and family life.

Interviews were usually conducted at the respondents' duty station, though on occasion commuting to a centralized location was required. Interviews were generally scheduled for 4-6 individuals, though, in practice they varied anywhere from 1 to 7. There were 5 interviewers in all, varying widely in number of interviews conducted, but all interviewing individually. Interviewers also varied as to sex, race, and military/civilian status. Interviews were usually 1 to 2 hours in duration, and were all tape recorded for later content analysis.

Results and Discussion

Several questions were asked about career involvement, officer role, and the Army lifestyle. There was a substantial amount of ambiguity about the meaning of the questions, both among respondents and among interviewers. The primary difficulty was the meaning of "career involvement." Did it mean attention on the job and career activities or did it mean intent to remain on active duty after the five-year obligation. In this analysis, responses which focused on the former alternative will be discussed in this section, while those addressing the five-year decision will be considered with a subsequent question on change in career commitment.

The most interesting results in responses to this question were in the differences between 1980 and 1981 graduates (Table 2). About 25% of the 1980 graduates indicated that they were more involved in their careers. The figure for 1981 graduates was almost 50%. There seems to have been a large shift in career interest during this one year time frame. The shift occurs for both males and females, although females are generally less likely to become more involved than males. A similar pattern is observed in the answers to questions on changes in adjustment. Class of '81 respondents are more likely to have experienced a positive change and males are more likely than females to have had a positive change. Again, interpretation of these results raises the question of statistical artifact or real difference. A significant proportion of the involvement change can be attributed to the radical change of males stationed overseas. Is this an accident or did the increased value of the dollar make everything seem rosier?

Answers to the questions on what factors influenced greater career involvement or adjustment to an officer's role shed little light on this issue (Table 3). Males in the class of '81 were much more likely to cite confidence, job satisfaction, and familiarity, but there seems to be no clue about why these factors should be more strongly felt on overseas assignments. They do, however, make sense as reasons for assuming a higher level of participation in an Army career.

Females from the class of '80 were more likely to express a negative change toward their officer roles than men. The majority of men who perceived a positive change simply felt more comfortable and familiar with their role. The majority of women who experienced a negative change were upset by different treatment for women and felt that men were just more likely to take the Army's "flack" than women. The latter comment probably says more about general attitude differences between male and female officers, as perceived by female respondents, than any other expressed in the interviews. Both groups perceive the same problems, for example both identify similar job dissatisfaction factors, but 1980 female graduates tend to be less tolerant of the negative aspects than men. Perhaps it is because many begin their careers from a more defensive perspective, as the first female West Point graduates. members of the class of '81 do not seem to have similar views.

Only a small number of individuals indicated a change in adjustment to the Army lifestyle. The most mentioned negative factors associated with the Army lifestyle are adjusting to time requirements and mandatory social functions. The most frequently cited positive factor is getting married. Six officers were married in the previous year and this seemed to significantly contribute to their overall adjustment or at least it made them happier.

NOTE: This document represents the views of the author and not the official position of the U.S. Army, or any other governmental agency unless so designated by other authorized documents.

The interview tapes were transcribed by Dr. Richard J. Orend under contract DAAG 60 85 M 1799. The research program was supported by contract 13 ARI 85-28 from the Army Research Institute (Jerome Adams, principal investigator).

REFERENCES

Adams, J. Report on the Integration of Women with the Corps of Cadets at West Point -- Project Athena IV. West Point, NY: AG Printing Office 1980.

Adams, J. & Yoder, J.D. Women Entering Non traditional Roles: Conflicts between Sex Roles and Work Proceedings: The Academy of Management National Convention. Boston, MA: 1984.

Adams, J. Project Proteus: Early Career Preparation, Experiences, and Commitment of Female and Male West Point Graduates. Vol I, II, & III. West Point, NY: AG Printing Office 1980.

Davenport, L.C. Dual-Career Couples: Spouses Influence on Career Decisions. Unpublished Master's thesis San Diego State University, Summer 1984.

Hall, D.T. Careers in Organizations. Santa Monica, CA: Goodyear 1976.

Hall, D.T. Organizational Commitment: Theory, Research, and Management. Unpublished paper Evanston, IL: Northwestern University, April 1979.

Hall, F.S. & Hall D.T. The Dual Career Couple. Reading, MA: Addison-Wesley 1979.

Moskos, C. "From Institution to Occupation: Trends in Military Organization" Armed Forces & Society 1977, 44-50.

TABLE 1
Total Participants

	Class of 80		Class of 81	
	Number	% of Total	Number	% of Total
Male	88	75.9	85	81.7
Female	28	24.1	19	18.3
Total:	116		104	
Stationed in CONUS*	83	71.6	80	76.9
Stationed OVERSEAS**	33	28.4	24	23.1
Total:	116		104	
CONUS Males	67	80.7	68	85.0
CONUS Females	16	19.3	12	15.0
Total:	83		80	
OVERSEAS Males	21	63.6	17	70.8
OVERSEAS Females	12	36.4	7	29.2
Total:	33		24	

*Including Hawaii

**Germany and Korea

TABLE 2
Change in Extent of Career Involvement

			Class of 80		Class of 81	
				Total Sample		
Change in Career Involvement:		More Involved	29	25.0	49	47.1
		Less Involved	15	12.9	9	8.7
				By Gender		
Change in Career Involvement:	Males -	More Involved	22	25.0	41	48.2
		Less Involved	8	9.1	6	7.1
	Females -	More Involved	7	25.0	8	42.1
		Less Involved	7	25.0	3	15.8
				By Location		
Change in Career Involvement:	CONUS -	More Involve	22	26.5	34	42.5
		Less Involved	8	9.6	8	10.0
	OVERSEAS -	More Involved	7	21.2	15	62.5
		Less Involved	7	21.2	1	4.2

TABLE 3
Factors Influencing Career Involvement

	Class of 80			Class of 81		
Positive - More Involvement	Males #	Females #	Total #	Males #	Females #	Total #
1. Confidence, job knowledge	2	0	2	10	1	11
2. Job satisfaction, enjoy work	5	0	5	13	2	15
3. Security	2	0	2	1	0	1
4. Willing to put in time	2	0	2	3	1	4
5. Competitiveness	1	0	1	0	0	0
6. Think about it more	0	1	1	3	0	3
7. Married to another officer	0	1	1	0	0	0
8. Comfortable	0	1	1	0	0	0
9. Sense of duty involved with goals	0	1	1	0	0	0
10. Increased responsibility	0	0	0	2	2	4
11. Thinking about change in assignment	0	0	0	8	1	9
12. More in control	0	0	0	1	0	1
Negative - Less Involvement						
1. Frustration ("brick walls")	1	0	1	2	0	2
2. Lack job freedom	1	0	1	0	0	0
3. Current assignment	1	2	3	0	1	0
4. No other competition in life	1	0	1	0	0	0
5. Problems leave individual numb	1	0	1	0	0	0
6. Not in MOS	0	2	2	0	0	0
7. Dissatisfied with commander	0	1	1	0	0	0
8. Not enough time with troops	0	3	3	0	0	0
9. Poor commanders	0	3	3	0	0	0
10. Can't quit/move	0	2	2	0	0	0
11. Not good career for women	0	3	3	0	1	1
12. Senior officers to me are money oriented	0	1	1	0	0	0
13. Not motivated to keep up in field	0	0	0	1	0	1
14. Not as concerned about always doing well	0	0	0	0	1	1
15. Future in Army seems dim	0	0	0	2	0	2
16. Disillusioned by WP on role of officer	0	0	0	1	0	1

Mentoring in the United States Air Force: The Mentor's Perspective

Captain Francis Lewandowski
9th Organizational Maintenance Squadron, Beale AFB, CA

Captain Benjamin L. Dilla
Air Force Institute of Technology, Wright-Patterson AFB, OH

As a follow-on to Roche's 1979 survey of business executives and Uecker's 1984 survey of Air Force officers exploring the mentoring concept, this study examined mentoring from the perspective of senior Air Force officers who have been both proteges and mentors. Respondents were 95 Air Force officers selected for the 1985 entering class at Air War College. Of these, 58% responded that they have had a mentor and 48% stated that they have been (or currently are) mentors for junior officers. Comparisons were made to the earlier data gathered by Roche and by Uecker with respect to military background, career factors, and effects of mentoring on the respondents. Furthermore, roles fulfilled by a mentor were contrasted from the perspectives of the protege versus the mentor. Results for this sample of senior officers verified the prevalence and perceived importance of mentoring in the United States Air Force.

Introduction

Mentoring has been defined as a relationship between a senior member and a junior member of an organization in which the senior member is influential in molding and shaping the career of the younger member (Uecker, 1984). The concept of mentoring has recently received considerable attention throughout the field of management. Trade journals, in particular, abound with articles ranging from cross-gender mentoring to reasons why one should (or should not) enter into a mentoring relationship.

Empirical work in the area is limited to two recent surveys. In a study by Heidrick and Struggles, Inc., and reported by Roche (1979), 1250 executives recently appointed to their positions were surveyed. This study found that "nearly two-thirds of the respondents reported having had a mentor or sponsor" (Roche, 1979, p. 14). The research also discovered that mentored executives earned more money at a younger age and had a higher degree of satisfaction with their jobs and their career progress. Furthermore, those in the mentored group were better educated and more likely to have formulated and followed a career plan.

A modified and expanded version of Roche's survey was recently applied in the U.S. Air Force; Uecker (1984) surveyed 252 officers attending Air Command and Staff College (ACSC) and Air War College (AWC) to examine the prevalence and effects of mentoring in the Air Force. Results of this survey were reported at last year's Military Testing Association conference (Uecker & Dilla, 1984). Mentoring was not found to be as prevalent in the military sample (42% versus Roche's 64%), but the effects were similar. Mentored officers were better educated and more likely to have formulated a career plan. They were more likely to have been promoted early (a parallel to Roche's finding of earning more money at an earlier age) and had greater job and career progress satisfaction (Uecker & Dilla, 1984).

An important extension of Roche's survey in Uecker's research involved examination of the roles or functions played by a mentor in helping to develop the subordinate. The survey asked respondents to categorize ten potential roles identified by Lea and Leibowitz (1983) as being primary, major, secondary roles, or roles not played by their mentors. Results showed that respondents saw their mentors primarily in the sense of role models, motivators, and advisors rather than in more directive roles such as sponsor and protector (Uecker & Dilla, 1984).

Since the prevalence of mentoring in the Air Force officer corps had been previously supported, the thrust of this project was to reexamine the prevalence of mentoring and to further investigate the phenomenon from the mentor's perspective. To accomplish this, a sample of high-potential senior officers was surveyed to find out if they had had mentors, if they had become mentors for others, and, in general, to examine their point of view concerning the mentoring process within the Air Force. The survey, based on those of Roche (1979) and Uecker (1984), also attempted to estimate the perceived effect the mentor had on the career of his protege and on the Air Force.

Methodology

Sample

The sample for this study needed to be drawn from a population of Air Force officers senior enough to have had the opportunity to be mentors as well as to have had mentors. Air Force policy precluded sending questionnaires to general officers who would most closely parallel Roche's (1979) sample in meeting this criteria. Authorization to survey designees for the 1985 entering class of AWC was granted. These officers, 112 in number, were lieutenant colonels and colonels (O-5 and O-6) coming out of a variety of leadership and staff positions including squadron commanders, directors at air division level, and system program directors. Furthermore, they had been identified, by the fact of their selection for AWC, as having high potential for further advancement.

Procedure

Surveys were mailed to officers at their duty addresses several months before their departure for AWC. Participation was voluntary, and respondents were assured of anonymity. They were asked to mark their responses directly on the survey instrument and return it in a postage-paid return envelope. Because of the frequent negative connotation of mentoring in the Air Force, the survey cover letter defined mentoring and carried an endorsement by the Dean of the School of Systems and Logistics (a USAF O-6).

Measures

The survey instrument was based on Uecker's (1984) questionnaire used the previous year with AWC and ACSC students. The format of the survey was expanded so that the officers would respond to items separately from the perspective of being a protege and from being a mentor. Eighteen items focused on the officer as protege, fourteen as a mentor; fourteen items were concerned with desirable characteristics of a mentor, and ten items addressed personal background, promotion history, and current satisfaction. Space was

provided to list important characteristics of a protege and other open-ended comments and suggestions at the end of the survey.

Results

Respondent Profile

Of the 112 officers sent surveys, a total of 95 (85%) responded. The majority received their commission from ROTC (58%) vice OTS (27%) or a service academy (15%). The median response for age at commissioning was 22 years, although ages ranged from 20 to 33. Most had an advanced degree (93%) and had received at least one below-the-promotion-zone (BPZ) promotion (87%). The largest group identified with no single major command (36%) although the three largest commands were well-represented--SAC (21%), TAC (16%), and MAC (12%).

Mentoring Experience

Of the 95 respondents, 58 reported having had a mentor who took a personal interest in them and guided or helped mold their careers. This 61% rate of mentoring is very close to the 64% reported by Roche (1979) for the private sector. Using a normal approximation to the binomial distribution and comparing the percentages, no significant difference was found. The median response for the number of mentors was two, with a range from one to six. Most (23 of the 58) said their mentor had first exhibited an interest in them only fairly recently, after the tenth year of service. The majority (35) indicated that they still had a relationship with their mentor, although most of these (24) described the relationship as "friendly" rather than "close". The largest group (26) reported having had a general officer as a mentor; the next largest group (12) said their mentor was their immediate supervisor. When asked how much influence their mentor had exerted, most said it was substantial (25) or moderate (21); only a few chose the extremes of extraordinary (5) or little (7) influence.

In the second section of the survey, 46 officers (48% of the sample) stated that they had served as a mentor to another individual. Two-thirds of these officers (30 of the 46) reported that they currently had one or two proteges. When asked how long their longest mentoring relationship had lasted, the modal response was a tie between two and three years. This finding may be an indication that these officers had only recently moved into positions of command and leadership where they could be mentors, or it may be a result of the very mobile Air Force way of life, such that relationships do not extend much beyond one assignment. When asked how much influence they had exerted over their proteges, again there was a fairly equal split between substantial (22) and moderate (24) influence with none reporting extraordinary or little influence.

Effects of Mentoring

This study failed to find any significant differences between mentored and unmentored groups with regard to formulation of a career plan, job satisfaction, or early promotions as had been found in the previous research. However, the small sample size of this study may have been a limiting factor. A significant difference was found between groups with respect to career progress satisfaction ($t=-2.85$; $p<.01$), with the mentored group reporting greater satisfaction.

For the 46 officers who had served as mentors, there were no significant differences with respect to early promotions or career progress satisfaction; however, they had significantly higher job satisfaction than their contemporaries who had not been mentors ($t=-2.25$; $p<.05$).

Roles of the Mentor

This study examined roles of the mentor from the perspective of both proteges and mentors. This allowed for a contrast between the two perspectives as well as a comparison with Uecker's (1984) results for proteges from a similar sample. For each of the ten roles identified by Lea and Leibowitz (1983), respondents were asked to indicate if the role was "most important" (assigned a scale value of 3), major (2), secondary (1), or a role not played (0). Mean responses for each role for this study's group of 46 mentors, 58 proteges, and Uecker's (1984) group of 106 proteges are presented in Table 1. Rankings of the ten roles within each group are also presented for ease of comparison.

Table 1
Roles of the Mentor

Role	Average Rating[1] (& Rank within Group) AWC Mentors (n=46)	AWC Proteges (n=58)	AWC & ACSC[2] Proteges (n=106)
Advisor	2.158 (1)	1.706 (5)	1.853 (2)
Counselor	1.932 (2)	1.708 (4)	1.598 (5)
Motivator	1.798 (3)	1.734 (2)	1.800 (3)
Role Model	1.711 (4)	1.840 (1)	1.924 (1)
Guide	1.675 (5)	1.321 (8)	1.500 (7)
Teacher	1.611 (6)	1.344 (7)	1.441 (8)
Communicator	1.561 (7)	1.093 (9)	1.505 (6)
Supporter	1.500 (8)	1.511 (6)	1.613 (4)
Sponsor	1.343 (9)	1.716 (3)	1.426 (9)
Protector	1.095 (10)	0.713 (10)	0.964 (10)

[1] Ratings and assigned scale values were:
Most Important = 3; Primary = 2; Secondary = 1 ;
Not Played = 0.

[2] Data adapted from Uecker, 1984.

The largest difference in ranks occurred for the controversial role of "sponsor". This term, which often carries a negative connotation in the Air Force, was rated relatively low by Uecker's (1984) sample and by the smaller group in this study which rated the roles from the perspective of being a mentor. Yet, when rated from the perspective of being a protege, the respondents of this study gave it the third highest mean rating; in fact, it received the largest percentage of "most important" responses for this sample.

Other notable differences included the advisor role, which received the highest mean rating for the group of mentors (also the highest absolute rating in any of the groups), but emerged fifth when rated from the protege's perspective in this sample (second in Uecker's [1984] sample). In the opposite direction were the results for "role model", rated highest by both groups of proteges but only fourth among the roles rated by the mentors.

Discussion

This study found mentoring in the Air Force to be as prevalent in the Air Force as in Roche's (1979) private sector study. Despite this similarity, there are distinctions in the nature of these relationships. In Roche's study, two-thirds of the respondents reported that a mentoring relationship began for them in the first five years of their career, while the majority of officers stated their relationships did not commence until after ten years of service. Also, relationships in the Air Force seem to be shorter and more varied. Only one-third of Roche's respondents reported having had two or more mentors (despite their earlier start), while 73% of the AWC officers (64% in Uecker's [1984] study) reported two or more mentors.

Roche (1979) found that 62% of the 1250 executives had proteges. In this study, only 48% of the officers indicated they had proteges; interestingly, most of these had previously had a mentor, so the phenomenon seems to be largely self-perpetuating.

With regard to the effects of mentoring, officers who had mentors had greater career progress satisfaction, while those who served as mentors had higher job satisfaction. These results are consistent with previous findings (Roche, 1979; Uecker, 1984) and with the effects predicted from career development theories discussed elsewhere (Dilla, 1985).

Contrary to Roche's (1979) findings, no difference was found for this sample with regard to formulation of a career plan. It should be noted that Uecker found an relationship for his total sample but not for the AWC respondents alone. This difference from Roche's results may be due to the centralized reassignment processing within the Air Force. Respondent comments indicated that "needs of the Air Force" and unforeseen career opportunities sometimes dictated career plan changes.

Regarding the roles of the mentor, proteges most often described their mentor as a role model, although the mentors assigned less importance to this function. This difference seems natural since it would be difficult for the mentor to tell to what extent the protege is observing and striving to emulate his behavior. The most important roles from the mentors' perspective were those of advisor, counselor, and motivator, roles that are more active but not highly directive. Proteges were in agreement with the relative importance of the motivator role but seemed less willing to admit that their mentors had to "counsel" them. This difference may be due to some of the negative connotations to the term itself.

There was clear agreement across all groups that the mentor does not serve as a protector to the protege, or at least does so very infrequently. It is noteworthy that the mentors assigned greater importance to this role than either group of proteges; further study of the occurrence of this function may be merited. Respondents also tended to play down the importance of the sponsor role, with the exception of the AWC proteges. It appears that the more senior group of AWC officers perceived that their mentors had provided growth opportunities for them to a greater extent than did Uecker's

(1984) combined sample of AWC and ACSC students. However, the low rating by the same group of officers when viewing the roles as mentors is puzzling. Further research with more detailed definitions of the terms and larger sample sizes may be the only way to resolve or clarify these differences.

Comments at the end of the surveys indicated that many officers understood and supported the use of mentoring in the Air Force; however, many harsh, negative comments revealed stereotypes and misconceptions even at this senior officer level. The Air Force should publicize the reasons for, and potential benefits of, the informal mentoring process. Furthermore, research on mentoring should continue and be expanded to both a broader scope and higher level of officers in order to better understand the dynamics and effects of the process.

References

Dilla, B.L. (1985). Mentoring: Cause and effect of good leadership. Proceedings of the Association of Human Resource Management and Organizational Behavior, pp. 630-635.

Lea, D., & Leibowitz, Z.B. (1983). A mentor: Would you know one if you saw one? Supervisory Management, 28, 32-35.

Roche, G.R. (1979). Much ado about mentors. Harvard Business Review, 57(1), 14-28.

Uecker, M.E. (1984). Mentoring and leadership development in the officer corps of the United States Air Force. Unpublished master's thesis, 84S-30, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Uecker, M.E., & Dilla, B.L. (1984). Mentoring as a leadership development tool in the United States Air Force. Proceedings of the 26th Annual Conference of the Military Testing Association, pp. 423-428.

Preparing to Evaluate Career Path Changes: Pre-Change Database Development

Robert E. Chatfield
Robert F. Morrison

Navy Personnel Research and Development Center

The purpose of this study was to establish, if feasible, a pre-change database that the Navy could use to assess the effectiveness of major revisions to the Surface Warfare Officer (SWO) career path.

Background

The Surface Warfare community has traditionally stressed that its officers be "generalists" with diversified work experience perceived to be career enhancing. However, recent program and career assignment policy changes have implemented changes that increase specialization prior to the Executive Officer (XO) tour. The origin of these significant changes to the Surface Warfare Officer (SWO) career path can be traced to the 1981 Surface Warfare Commanders Conference (SWCC) where the issue of fleet readiness and its relationship to SWO technical competence and experience was discussed. The SWCC concluded that the generalist approach to officer development actually undermined and degraded readiness by not allowing officers to gain the requisite technical experience required to manage/operate today's complex shipboard systems. The following FY81 statistics were cited to support their contention:

--Only 41% of engineering department heads had prior engineering experience as a division officer.

--Only 38% of operations department heads had prior operations experience as a division officer.

--Only 48% of weapons/combat systems department heads had prior weapons/combat systems as a division officer.

In every instance, fewer than half of the officers had <u>any</u> work experience in the type of department that they were being asked to manage.

As a direct result of the SWCC concern, major revisions were made to SWO training programs and career policies over the next two year period. The goal of the revisions, promulgated as NAVOP 105/83, was to develop more technically experienced officers at the key department head level, which would in turn be expected to promote an increase in the operational readiness of ships. As a department head, the SWO must manage the maintenance, operation, and employment of complex systems/platforms.

The new career path is tailored to provide them with the opportunity to gain specialized technical expertise in the first 10-12 years of their Navy career without the expectation of being a generalist. Specifically, the new career path is structured for specialization in one of the three major surface warfare areas: engineering, operations, and weapons/combat systems.

Advocates of the revised SWO career path insist that this shift toward specialization was technology-driven and inevitable. However, critics argue that these newly created "specialists" will encounter difficulties at the XO/CO level where familiarity with the entire ship is critical. Only time, and a well planned evaluation of the impact of the revisions will prove which of these contrasting views is indeed correct.

Objectives

The overall objectives of the study were:

1. Identify existing measures of ship department readiness and performance.

2. Build a pre-change (prior to implementation of the new career path) database consisting of FY82-FY84 data for the most promising measures.

3. Assess the consistency across measures and the stability of individual measures over the three year time period.

4. Identify measures acceptable for inclusion in a pre-change database that the Navy could use 3-5 years hence to evaluate the impact of the career path revisions on fleet readiness.

As the project was of short duration (9 months) and funding was limited, it was not possible to design and implement procedures for evaluating career path effects. Therefore, the study was limited to investigating the appropriateness of using existing measures of ship department readiness and performance to assess the impact of the changes.

Method

A multiple measure, or "performance profile," approach was taken under the assumption that no single performance measure could acceptably serve as an evaluative standard for the new career path. It was theorized that by focusing on essentially a battery of readiness indices, the impact of the new policies could be inferred. An analysis of inter-measure consistency and year-to-year stability of individual measures would indicate if the multiple measure approach was appropriate.

Two different classes of data were used to build a preliminary pre-change database. The first class, yielding "ship data," involved identifying existing measures of departmental

readiness and performance and collecting data for those satisfying acceptability criteria. The second class provided "personnel data" that might describe the performance of department heads as reflected in the personnel records of individual officers.

In order to determine which ship data measures of readiness were appropriate for inclusion in the pre-change database, and to provide a means of cross-measure comparison, nine evaluative criteria were developed. For example, numeric data were required for quantitative summarization. Additionally, scores could not be inflated such that improvement potential and variability suffered. In general, measures had to be objectively scored, focus on department head performance, be available in a format facilitating rapid computer analysis, and be well accepted by the SWO community. These criteria were developed to address issues of reliability and validity while also reflecting project-related constraints.

Over fifty interviews were conducted with senior level officers in the Surface Warfare community to identify existing measures of departmental readiness. Information collected about measures included: measurement cycle; scoring procedures; fleet reputation, etc. The interviews yielded 21 "candidate" measures for possible inclusion in the pre-change database. These measures were then individually scored on the nine evaluative criteria. A measure was eliminated from consideration if its standing on any criterion was completely unacceptable.

Six of the 21 measures proved to be acceptable on all criteria, however, data for only three were actually collected and analyzed. The three remaining measures were not included because a substantial amount of clerical effort would have been required to extract relevant information and transcribe it to a format facilitating computer analysis. The three ship data measures included in the pre-change database were:

1. Propulsion Examining Board (PEB) Assessments (both the Lighting Off Examination (LOE) and Operational Propulsion Plant Examination (OPPE).

2. Nuclear Weapons Technical Inspections (Nuclear Weapons Acceptance Inspection (NWAI); Nuclear Technical Proficiency Inspection (NTPI); and Defense Nuclear Security Inspection (DNSI).

3. Departmental Excellence Awards (presented in conjunction with the Battle Efficiency Awards cycle).

Two types of personnel data were collected: 1) The proportion of department heads failing to complete their full tour during the FY82-FY84 period; and 2) The annual number of department heads involved in a Detachment for Cause (DFC) action. The former is often informally indicative of poor job performance while the latter is the formal administrative action for relieving an officer of his duties. The assumption is that the incidence of both of these occurrences would decrease if the new career path was having its intended effect.

Results

Ship data were evaluated by several different methods to determine if their inclusion in the pre-change database was appropriate. First, each measure was examined for year-to-year consistency. Substantial fluctuation or variability in pass rates across the three year time period was indicative of poor consistency and "reliability." While perfect consistency from year-to-year was not expected, 10%-15% variation was the maximum acceptable. Second, measures were assessed for trends in rising or falling pass rates across the three year time period. Finally, cross-measure comparisons were made to evaluate the extent to which the various assessments of departmental readiness were in agreement.

The generally poor stability of ship data performance measures is reflected in Figure 1 which presents LOE and OPPE pass rates for several ship types. Substantial year-to-year fluctuations in pass rate percentages are evidenced in the variability columns. For the LOE, variability was unacceptably high for both ship types presented. That is, yearly pass rate differences ranging from 20%-50% do not provide the stable performance baseline required to evaluate effects of the new career path. While OPPE variability was marginally acceptable for DD/DDG (11%/16%), it was completely unacceptable for LSD/LST (50%/89%). No consistent trends in pass rates were evident for either measure. Pass rates appeared to rise and fall in a random, unsystematic manner.

FIGURE 1

Examples of LOE and OPPE Pass Rate Instability

	PACIFIC			ATLANTIC			VARIABILITY (HIGH-LOW)	
SHIP TYPE	1982	1983	1984	1983	1984	1985	PAC	LANT
LOE Exam								
FF/FFG	100	67	88	100	88	80	33	20
LHA/LPH/LKA/LPD	86	60	60	33	50	0	26	50
OPPE Exam								
DD/DDG	76	65	68	75	71	88	11	16
LSD/LST	100	67	50	100	50	13	50	87

All numbers expressed as percentages.

In contrast to PEB inspections, NWTIs displayed rather good stability. A summary of NWTI pass rates for four ship types is provided in Figure 2. Several of the variability percentages fall below the 10%-15% tolerance previously discussed. Pass rates were exceptionally stable for Auxilliary class ships where the variability was only 4% in the Pacific fleet and 7% in the Atlantic fleet across the three year period.

FIGURE 2

Examples of NWTI Pass Rate Stability

SHIP TYPE	PACIFIC 1982	PACIFIC 1983	PACIFIC 1984	ATLANTIC 1983	ATLANTIC 1984	ATLANTIC 1985	VARIABILITY (HIGH-LOW) PAC	VARIABILITY (HIGH-LOW) LANT
CG/CGN	83	100	100	89	100	89	17	11
DD/DDG	85	88	74	85	81	84	14	4
FF/FFG	81	92	91	93	90	75	11	18
AUXILLIARY	72	71	75	81	74	79	4	7

All numbers expressed as percentages.

Inter-measure comparisons made on a year-to-year basis indicated that there was little consistency among readiness indices. For example, there was essentially no relationship between ship types' pass rates on the OPPE and on the NWTI. Although all measures ostensibly assess the operational readiness of ship departments, inter-measure consistency was lacking.

The final ship data measure, departmental excellence awards, displayed both excellent (less than 5% variability) and poor (more than 35% variability) stability. There appeared to be some type of interaction between ship type and award category as awards were often stable for one ship type but highly volatile for another. Because of this inconsistency, departmental excellence awards do not provide the stable performance baseline desired.

Analysis of personnel data measures revealed that neither was appropriate for inclusion in such a pre-change database. The percentage of department heads leaving the job early was quite small (maximum of 10.1%) and no year-to-year trends were evident. Figure 3 provides a summary of performance-related DFCs by major department for the three year period. Once again, very few department heads are represented in this measure as there were a total of only 59 such actions. Variability of DFCs year-by-year was greater than desirable and no trends were identified. For example, while DFCs were steadily on the rise in the Atlantic fleet, a similar trend was not evident for the Pacific fleet.

FIGURE 3

Performance-Related DFC Actions for Major Department Heads in FY82-FY84

	ENG		OPS		W/CS		TOTAL	
YEAR	PAC	LANT	PAC	LANT	PAC	LANT	PAC	LANT
1982	5	3	0	2	3	0	8	5
1983	4	7	4	5	2	0	10	12
1984	1	12	3	2	3	3	7	17
TOTAL	10	22	7	9	8	3	25	34

Conclusions

Analysis of ship data performance measures indicated that such indices were not stable enough to establish a performance baseline for assessment of career path influences. The NWTI was the only measure to display any promise for such an application. In addition, consistency among multiple measures of readiness and performance was not evident. Results from one type of evaluation were not in agreement with those from another on a year-by-year basis.

Personnel data measures proved to be equally unacceptable for evaluating the impact of the new career path on readiness. In both instances, the number of department heads represented was very small and stability was lacking.

In conducting analysis of these existing measures of departmental readiness and performance it became evident that there are a variety of factors beyond the scope of the department heads' control that can moderate results. While a number of situational (e.g., experience level of department personnel, CO influence) factors are essentially random and other miscellaneous factors (e.g., age of ship) can be statistically controlled for in analysis, there remain a plethora of systemic factors whose influence cannot readily be identified. For example, such factors as the fleet's operations tempo, changing evaluation standards, budgetary limitations, etc. can dramatically affect readiness assessments. The influence of these systemic factors undoubtedly contributed to the poor stability of individual measures and lack of inter-measure consistency.

In summary, multiple measures of departmental readiness and performance evaluated in the present study should not be used to assess the revised SWO career path. In light of the poor stability documented for existing measures of ship department readiness and performance, it is recommended that the Navy conduct a critical review of these measures in an effort to improve their reliability and validity.

USAF OFFICER PROFESSIONAL MILITARY EDUCATION[1]

by

JOHN M. BELL, First Lieutenant, USAF
Directorate of Military Occupational Structures
National Defence Headquarters, Ottawa, Ontario, Canada

and

JOSEPH S. TARTELL
Occupational Analysis Division
USAF Occupational Measurement Center, Randolph AFB, Texas

INTRODUCTION

To ensure that USAF officers will have the skills to carry out their leadership and managerial responsibilities, the Air Force provides a variety of precomissioning and postcommissioning Professional Military Education (PME) courses that can be taken at specific career points. To determine whether these courses are truly responsive to the needs of USAF personnel, the Air War College requested that the USAF Occupational Measurement Center (USAFOMC) conduct an occupational survey that would help validate or redesign the curricula of officer PME courses. Specifically, USAFOMC was asked to determine leadership, management, and communicative tasks performed by company and field grade officers, and to determine the need of, or benefit from, the various PME schools and courses. Precommissioning PME sources are the USAF Academy, AFROTC units, and Officer Training School (OTS). Postcommissioning PME sources include Squadron Officer School (SOS) for junior officers, Air Command and Staff College (ACSC) for middle level officers, and Air War College (AWC) for senior officers.

METHODOLOGY

To meet the request, USAFOMC developed and used five separate survey instruments to collect data. The first of these was a task list containing 347 leadership, management, and communicative tasks under 14 duty headings. These tasks were to be rated by survey respondents on a 9-point scale according to the relative amount of time spent on each task, compared to the time spent on each of the other leadership, management, and communicative tasks they performed.

Second, data on the difficulty of the same leadership, management, and communicative tasks discussed above were collected via a task difficulty booklet. Difficulty was defined as "the amount of time needed to learn to do each task satisfactorily." Respondents were asked to rate each task on a 9-point scale according to its relative difficulty, compared to the other tasks.

Third, using an inventory with the same leadership, management, and communicative tasks discussed above, respondents were asked to rate each task on a 10-point scale according to its need in Air Force educational programs.

A fourth set of data was collected via a survey containing a list of 275 topics from the curricula of officer PME courses. For each of these topics, respondents were asked to rate the extent to which knowledge of, or skill in, each topic was necessary to perform their present job (need-in-job).

Using the same list of topics, respondents to a fifth booklet were asked to rate the extent to which knowledge of, or skill in, each topic was necessary to function as a career officer (need-in-career). An 8-point scale was used to rate the topics in both of these surveys.

These surveys were validated and approved by representatives of the various PME schools. Random samples were selected for administration of the surveys between June 1983 and April 1984. Representative samples across the surveys were achieved.[2]

RESULTS

The analysis of the task performance data showed a pattern of increasing involvement in leadership, management, and communicative tasks as officers increased in rank from lieutenant to colonel. Supporting this pattern were data that showed the percentages of officers who had supervisory responsibilities increased from 38 percent among lieutenants to 95 percent among colonels. Additionally, the percentage of total job time spent on the tasks in the survey increased from 56 percent to 81 percent, from lieutenant to colonel, respectively. Related to these was the organizational assignment pattern, which showed the manner in which the percentage of officers assigned to organizational levels as rank increased. This pattern of increasing involvement is not surprising, but it does illustrate the changing nature of most officers' responsibilities. Further, it provides some rationale for a continuing multiphased professional development program.

The data showed a great amount of diversity in the tasks performed across career fields and ranks.[3] At the same time, there was a substantial amount of similarity in certain tasks performed, particularly in some tasks dealing with communication skills and motivating others. Further, the differences between ranks in relative time spent on tasks in each duty were very small.

The analysis of the education emphasis data revealed a low reliability of raters. In short, insufficient agreement on the amount of education required to perform tasks existed among officers in general and across career fields and ranks. This finding reemphasized the diversity of opinions and needs concerning USAF officer PME.

The reliabilities of task difficulty data were quite high.[4] Fifty-six tasks received high difficulty ratings from the total group of raters, and most were communicative tasks, such as drafting or writing relatively high level documents (officer effectiveness reports, plans, staff papers, and reports). Other highly rated tasks involved skills such as determining resources; administering disciplinary actions to civilians; and ordering, persuading, or influencing those superior in rank or position. Fifty-eight tasks received low difficulty ratings. Many of these, also, dealt with communicating and motivating, but were of much lower-level activities, such as drafting or writing short note replies and reading professional publications. Other lowly rated tasks include providing informal feedback, attending training sessions, and maintaining appearance standards.

Analysis of officers' self-perceived need of various PME topics in their jobs and in their careers showed a great deal of diversity within most of the career fields and ranks. In spite of this diversity, it was possible to create a rank order listing of topics from each of these groups displaying the relative need of these topics in the job or in the career. In general, the data showed those PME topics which officers believed were needed in their jobs were topics they also needed in their careers; conversely, they generally felt those topics not needed in their job were not needed in their careers. Only minor differences were seen in perceptions across these two studies. Communication topics (eg, effective listening, active writing, logical thinking) were rated as most needed by officers in their jobs and careers; senior officers saw the needs as about equal, while junior officers perceived them as greater in their job than over their career. Topics on the military environment, national security, and military employment generally received the lowest relative need ratings, both in terms of in the present job and over the career. Rated lowest were topics on other services policies and doctrines, economic theories and systems, and foreign relations.[5]

Background data, collected in all five surveys, were used to assess, among other things, officers' job satisfaction and perceptions of benefits from PME. Of the 10,607 responses to the 5 parts of the project, 10,177 were used in assessing background data responses. (The difference here reflects the elimination of duplicate responses of officers who were asked to complete more than one kind of survey booklet, since background questions across booklets were identical.)

While the job satisfaction indicators were high, perceptions of benefits from USAF PME were, at best, mixed. Those who participated in precommissioning PME at the USAF Academy, indicated the highest degree of benefit, while the extent of benefit from ROTC and OTS PME was much lower. Officers indicated very low benefit from SOS by correspondence, and lower benefits in general from PME by correspondence or seminar, than through residence programs.

COMPARISON TO PREVIOUS RESEARCH

Results of this survey were compared to those of the Officer Professional Military Education Curriculum Validation Project produced by USAFOMC in August 1980 and the Survey of Air Force Officer Management Activities and Evaluation of Professional Military Education Requirements produced by the Air Force Human Resources Laboratory in December 1969.

While the scope and complexity of PME studies increased over the years, certain results were consistent. Analyses of task involvement over time, for example, showed a consistent increase in the level and number of leadership, management, and communicative tasks with every increase in rank. Differences in task performance across utilization fields were also observed across studies; operations personnel, for example, were consistently lower performers (compared to other fields) at lower ranks, increasing the scope of their leadership, management, and communicative responsibilities as they increased in rank to the point of equality with other fields at the grade of colonel.

Task difficulty data (collected in this and the 1980 studies) generally agreed on the tasks officers consider most and least difficult. Heading the lists were drafting or writing high-level official correspondence, conducting high-level investigations, and determining resources. Low on both lists were attending training sessions, maintaining personal appearance standards, and drafting or writing low-level correspondence.

The perceptions of need of particular curriculum topics were analyzed differently in this study than that data in the two previous surveys, but two similar findings were evident. First, there was a great amount of diversity within sub-groups analyzed as to what PME topics officers perceived they needed. Second, there was a general overall consistency as to what were the most and least needed topics within sub-groups. While previous studies used average ratings and this one used a rank order method to assess need, agreements on self-perceived needs across time were striking: topics dealing with oral and written communication, leadership, and principles of management were consistently viewed as among the most needed, while those dealing with Army and Navy doctrine, international relations, and warfare were perceived as least needed.

A comparison of the perception of benefits from the various methods of PME between the two most recent studies showed few differences. Both studies found the perception of benefits from SOS by correspondence to be small, from residence programs to be greather than for correspondence or seminar programs, and from USAF Academy PME to be of greater benefit than ROTC or OTS-OCS PME.

CONCLUSIONS

Data from this project provide a good view of what leadership, management, and communicative tasks officers perform; the relative difficulty of tasks in the judgement of Air Force officers; and rank order listings of curriculum topics which officers perceive they need in their jobs and their careers. Based on the data, several conclusions about Air Force officer PME may be drawn.

First, there are differences in the leadership, management, and communicative task performance of USAF officers across career fields and ranks. Further, there is a lack of agreement among officers as to what PME topics are required in their jobs and careers.

Second, there is a progression in terms of the number and complexity of tasks performed by officers as they advance in rank. Based on this, there is sufficient rationale for the continuation of a multiphased PME program, although it is doubtful a program designed purely on differences in rank can be of equal value to all officers.

Third, the success of a PME program must be judged on its usefulness to the most people. While concentrating on the PME-related tasks most officers at any given phase perform, PME must also educate officers as to the need and importance of a broader understanding of military-related topics to their roles as professionals.

NOTES

1. This paper is based on Professional Military Education -- Officer, AFPT 90-XXX-522 (USAF Occupational Measurement Center, October 1984). Copies of the report are available upon request from USAFOMC/OMY, Randolph AFB, Texas 78150-5000.

2. The return rates for these inventories was as follows: Task List - 2,016 usable/3,638 administered: 55 percent return rate; Task Difficulty - 312/584: Education Emphasis - 316/598: 53 percent; Need-in-Job - 4,100/6,742: 61 percent; and Need-in-Career - 3,863/6,384: 61 percent.

3. The career fields, in descending order of the average number of LMC tasks performed (and the number) were: logistics (146); legal (141); security police (140); personnel resources (139); chaplains (132); comptroller (119); scientific and engineering (102); cartography, geodesy, and intelligence (102); and operations (95).

4. A series of related computer programs, called the Comprehensive Occupational Data Analysis Programs (CODAP), was applied to the data to assist in the analysis. Reliability of individual raters was .30 and reliability of raters as a group was .99, both of which were well within the range of reliabilities accepted in USAF occupational research.

5. Correlations of rankings of need-in-job by rank ranged from a high of .99 between majors and lieutenant colonels to a low of .91 between lieutenants and colonels. Lowest correlations were between rankings by direct-commission officers and academy graduates (.75) and between cartography-geodesy-intelligence specialties and other utilization fields. Correlations of rankings of need-in-career by ranks, commissioning sources, and utilization fields were slightly lower, but comparable.

Visual, Auditory, and Reaction Time Measurement of Speed of Information Processing and Intelligence

Dennis P. Saccuzzo[1]
San Diego State University

and

Gerald E. Larson and Bernard Rimland[2]
Navy Personnel Research and Development Center, San Diego

Abstract

A battery of tachistoscopic, auditory, reaction time and microcomputer generated measures of speed of information processing were administered to 96 college students between 18 and 22 years of age. In addition, each subject was given a battery of tests designed to provide independent evaluations of right and left hemispheric functioning. Criterion measures included a verbal (Vocabulary) and nonverbal (Block Design) measure of intelligence. Results revealed a general processing speed factor in addition to task specific sources of variability. Morover, the processing speed tasks loaded on the same second-order factor as did the traditional measures of intelligence and aptitude. The findings support the theoretical view that processing speed may be a general factor in individual differences in performance on complex intellectual tasks. An important objective for future work in this area is to separate and evaluate the common and specific sources of variability on processing speed tasks, such as those used in this study, which contain little or no intellectual content and involve little or no complex problem solving.

[1]Reprint requests to the Psychology Clinic, San Diego State University, San Diego, California 92182. Funding for this work was provided by a faculty fellowship from the American Society for Engineering Education and by Future Technologies for Manpower and Personnel (PE62763, RF63-521-806) *. The authors wish to express their appreciation to Steven M. Paul for his assistance in the statistical analyses.

[2]Now Retired.

* The views expressed in this paper are those of the authors, are not official, and do not necessarily reflect the views of the Navy Department.

Paper and pencil tests of ability reached an asymptote more than 60 years ago in predicting job performance. Consequently, there has been a search for new approaches to the measurement of individual differences in ability and intelligence. One such approach, referred to as chronometric measurement, attempts to evaluate an individual's speed of information processing. The potential of chronometric measurement in military settings, however, is unclear due to a limited data base and some seemingly conflicting evidence. Though the preponderance of studies have revealed a small but significant correlation between processing speed tasks and those involving complex cognitive behavior, the meaning of such findings is a matter of controversy (see Hunt, 1980; Vernon, 1981).

In the present study, we obtained visual, auditory, and reaction time measures of processing speed for the same subjects. We also explored the feasibility of using microcomputer controlled tasks to evaluate processing speed, as suggested by Cory, Rimland, and Bryson (1977). Finally, we attempted to probe more deeply into the meaning of performance on a processing speed task, especially as it might relate to the functioning of the right and left cerebral hemispheres.

Method

The subjects were 96 San Diego State University students from an introductory course in psychology. All were between 18 and 22 years of age. Fifty-three were male; forty-three female. Eighty were Caucasian; the remaining 16 were black, Hispanic, and Asian. The sample was representative of the total population of San Diego State students in the 18-22 age range.

Each subject was tested on a set visual, auditory, and reaction time processing speed tasks. Subjects were also given the Vocabulary and Block Design subtests of the Wechsler Adult Intelligence Scale, Revised (WAIS-R) and the Cognitive Laterality Battery.

Two types of visual tasks were used, with the order of presentation alternating between subjects. Those of the first type involved tachistoscopic presentations in a paradigm that followed the general procedures used by Saccuzzo et al (1979). A microcomputer presented battery of five tests for estimating visual processing speed, as described by Larson and Rimland (1984), provided the second set of experimental tasks. Each microcomputer task was presented twice to each subject, once via a TRS-80 and once via an Apple II, in a counterbalanced design. Auditory speed of processing was measured by the Repetition Test of Tallal and Piercy (1973). A reaction time task, as described by Larson and Rimland (1984), was presented on the TRS-80 computer and keyboard. Each subject's median reaction time (based on 11 trials) was used as the index for the one (RT1), three (RT2), and five (RT3) choice tasks. The Cognitive Laterality Battery (CLB), developed by Gordon (1983) to evaluate individual differences in hemispheric asymmetries, was administered to subjects in small groups (8-16 subjects) after all processing tasks had been completed.

Results

Table 1 presents the results of a Schmid-Leiman hierarchical factor analysis in which all factors are orthogonalized and first-order factors are residualized (i.e., their correlated variance is "absorbed" into the second-order general factor). As the analysis shows, the processing speed tasks and the criterion variables loaded together on a second-order factor, which was labeled General Mental Speed. Four first-order factors, Reaction Time, Auditory Speed, Psychometric Intelligence, and Visual Speed, also emerged.

Table 1

Hierarchical Factor Analysis for Criterion Variables and Measures of Processing Speed

	Factor 1 (Second order) General Mental Speed	Factor 2 Reaction Time	Factor 3 Auditory Speed	Factor 4 Psychometric Intelligence	Factor 5 Visual Speed
AUDLG	.63	.07	.64	-.02	-.01
AUDST	.58	-.07	.67	.02	.01
ISIc1	.23	-.03	-.07	.02	.32
ISIc2	.22	.01	-.01	-.11	.38
TRAV	.47	.09	-.01	.11	.33
APAV	.40	-.08	.09	-.03	.43
RT1	.52	.61	-.02	-.03	.03
RT2	.61	.68	.08	.00	-.06
RT3	.52	.61	-.06	.03	.03
HSGPA	.30	-.01	-.07	.56	-.07
FRGPA	.40	-.15	.12	.44	.06
SAT	.47	.08	.06	.56	-.11
VOCAB	.32	-.12	-.07	.60	.02
BD	.50	.19	-.05	.37	.10
Total Variance	21%	9%	7%	10%	4%

* Salient factor loadings underlined. Correlations in the original matrix were reflected so that good performance has been positively correlated with all other variables.

AUDLG and AUDST = Auditory Processing Speed for long and short interstimulus intervals, respectively. ISIc1 and ISIc2 = The Critical Interstimulus Interval, trials one and two, respectively (a tachistoscopically determined nonverbal measure of speed of information processing). TRAV and APAV = TRS-80 and Apple II microcomputer derived scores for speed of visual processing. RT1, RT2, and RT3 = One, Three, and Five choice reaction times. HSGPA and FRGPA = High School and Freshman Grade Point Average. SAT = Scholastic Aptitude Test (total score). VOCAB and BD = Scores on the Vocabulary and Block Design Subtests of the Wechsler Adult Intelligence Scale, Revised.

Table 2

Correlations and First Unrotated Factor for Measures of Processing Speed and Hemispheric Functioning+

	Left Hemisphere	Right Hemisphere	Factor Loadings ++
AUDLG	.25**	.04	<u>.66</u>
AUDST	.21*	.07	<u>.54</u>
ISIc1	.16	-.19*	.03
ISIc2	-.03	.10	.21
TRAV	-.04	.22*	<u>.45</u>
APAV	.05	.23*	<u>.36</u>
RT 1	.01	.24**	<u>.70</u>
RT 2	.15	.07	<u>.77</u>
RT 3	.14	.08	<u>.65</u>
Factor Loadings ++	<u>.26</u>	<u>.30</u>	
Mean	.095	-.034	
SD	.56	.52	

+ Correlations have been reflected.

* $p < .05$

** $p < .01$

++ Factor loadings for first unrotated factor matrix for Principal Components factor analysis (total variance accounted for equals 25.4 percent). Salient loadings are underlined.

Right and Left hemisphere functioning was measured by the Gordon Cognitive Laterality Battery. AUDLG and AUDST = Auditory Processing Speed for long and short interstimulus intervals, respectively. ISIc1 and ISIc2 = The Critical Interstimulus Interval, trials one and two, respectively (a tachistoscopically determined nonverbal measure of speed of information processing). TRAV and APAV = TRS-80 and Apple-II microcomputer derived scores for speed of visual processing. RT1, RT2, and RT3 = One, three, and five choice reaction times.

Table 2 presents the correlations and first unrotated factor for the measures of processing speed and those for left and right hemispheric functioning. Examination of the table reveals a different pattern of correlations between the processing speed measures and the measures of hemispheric functioning. Specifically, auditory processing was related more strongly to the left hemisphere, visual processing to the right.

Discussion

A set of processing speed tasks loaded on the same second-order factor, derived from a hierarchical analysis, as did a set of traditional measures of intelligence and aptitude. This finding reveals that performance on complex cognitive tasks such as Block Design and the SAT is related to performance on tasks that have little or no knowledge content and require no complex problem solving strategy. The findings thus support the theoretical view that processing speed may be a general factor in individual differences in performance on complex intellectual tasks.

The data further reveal two major components to the variability in measures of speed of processing: a general component that accounts for roughly half the variance and task-specific components that account for the other half. These task specific components account for a fairly substantial share of the variability and cannot be ignored when considering the relationship between measures of processing speed and measures involving complex cognitive processing. The task specific variance, moreover, may be of intrinsic interest in that any given task may be related to important processes. For instance, the pattern of correlations revealed that the visual tasks tended to be related to the right hemisphere-related tasks on the Gordon Battery; the auditory tasks, by contrast, were related to the Gordon Battery left-hemisphere tasks. Hypothetically, it might be feasible to construct a set of processing speed tasks both to measure general intelligence, through obtaining a composite score in which individual differences due to task-specific abilities average out, and to measure more specific group factors that are relatively independent of each other and can help predict job performance. Furture research in this area should be directed toward separating the common variance from the task-specific and to determine more precisely what each of these components measures. We conclude that:

(1) Processing speed tasks, which contain little or no intellectual content and involve little or no complex problem solving skills, share common variance with conventional psychometric tests that do involve complex reasoning and problem solving skill.

(2) Scores on processing speed tasks are multi-dimensional. Though a general processing speed factor emerges from a hierarchical analysis of a set of visual, auditory, and reaction time tasks, more specific factors emerge as well. Consequently, it appears to be an oversimplification to ask whether any particular processing speed task is related to intelligence. Rather, more specific questions are needed.

(3) Measurement of Spearman's "**g**" factor of mental ability by means of speed of processing tasks will most likely depend on using a battery of such tasks having sufficient diversity in specific task features to permit the "averaging out" of task specific variance, so that the composite score will prdominately reflect the general ability factor that is common to all of the tasks.

References

Cory, C. H., Rimland, B., & Bryson, R. H. (1977). Using computerized tests to measure new dimensions of abilities: An exploratory study. Applied Psychological Measurement, 1, 101-110.

Gordon, H. W. (1983). Cognitive Laterality Battery. Western Psychiatric Institute and Clinic, University of Pittsburg, School of Medicine.

Hunt, E. (1980). Intelligence as an information-processing concept. British Journal of Psychology, 71, 449-474.

Larson, G. E., & Rimland, B. (1984). Cognitive speed and performance in basic electricity and electronics (BE & E) school. NPRDC TR85-3, October.

Saccuzzo, D. P., Kerr, M., Marcus, A., & Brown, R. (1979). Visual information-processing in mental retardation. Journal of Abnormal Psychology, 88, 341-345.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53-61.

Tallal, P., & Piercy, M. (1973). The Repetition Test. Neuropsychologia, 11, 389-398.

Vernon, P. A. (1981). Reaction time and intelligence in the mentally retarded. Intelligence, 5, 345-355.

Wechsler, D. (1981). Manual for Wechsler Adult Intelligence Scale-Revised. San Diego: Harcourt, Brace, & Janovich.

REPRESENTATION AND MODIFICATION OF HUMAN PERFORMANCE FACTORS IN TARGET SELECTION

Dwight J. Goehring
U.S. Army Research Institute Field Unit
Presidio of Monterey, California

I. Abstraction is Fundamental in Science

Abstracting or modelling of phenomena is a primary method of science for increasing knowledge. This approach is vital where the phenomenon under investigation cannot be subjected to direct scientific manipulation, such as astronomical processes or force-on-force combat. However, the quality of conclusions Obtained from a particular abstraction, or model, is critically dependent upon how completely and accurately the essential components and their interrelationships are represented.

Models which simulate the process of combat have long been important to military planners in the formulation of policy, strategy, and tactics (Brewer and Shubik, 1979). The use of modern computers has enabled the representation of an unprecedented level of detail in combat modelling. For example, extant Army combat simulation models consist of hundreds of thousands of lines of high-level-language computer code. This code represents the combat process in extreme detail, down to the field of vision and magnification specifications of the specific optics selected by every relevant weapon system at a given point in time.

Despite such great detail incorporated in many models, larger questions remain about the overall adequacy of the representation of the combat process. Characteristics of optical or weapon systems can be precisely determined using experimental methodologies. Data from such methods serve to determine with high confidence the values of many of the very large number of parameters present in modern combat simulation models. One class of phenomena which is not, however, so easily characterized is human behavior. As human performance factors (HPF) are an integral part of virtually all systems comprising the combat process, the nature of the representation of these factors in combat models has direct implications for the validity of any results.

II. Human Performance Factors in Models

There are several issues pertaining to the representation of HPF in the modelling of combat. Some of these can be illustrated by taking a simple behavior such as the likelihood that a single soldier will detect a particular type of vehicle under some situation specifying distance, atmospheric conditions, lighting, the soldier's physical and psychological states, and so forth.

One issue concerns the accuracy of estimating the likelihood of detection. With conventional experimentation and statistical methodology, an estimate of known certainty can be obtained. Similar methods can yield information on the variability between individual soldiers in detection likelihood. However, a serious difficulty arises with this approach when the detection parameters are sought for different sets of situation contexts. Human performance can vary not only as a function of each of the myriad of single factors but also based upon any possible interaction of the factors. Both combinatorial explosion and experimental impracticality become problems immediately. Thus, to rigorously establish model parameters of relevant contexts for HPF of even a simple task is simply not possible. Therefore, (1) HPF must be represented in combat simulation models in simplified form, and (2) at some point assumptions about such aspects as the value limits on HPF variables and the nature of interactions among variables become necessary.

When examples of representations of HPF in combat simulation models are considered, both the form and detail are found to vary widely. Miller and Bonder (1982) conducted an investigation of the HPF treatment in nine combat simulation models. Fifteen combat

processes (e.g. communications, maneuver control, and mobility, countermobility and survivability) were identified which contained 112 "human factor process phenomena" (e.g. decisions to request fire support, route selection of forces moving on the ground, increase in exposure due to encountering a minefield or obstacles). They observed that a phenomenon could be represented in a model in one of five basic forms: (1) assuming that the phenomenon is unvarying and always correctly performed, (2) representing the factor by a human player in real time while the model is running, (3) directly inputing the effect of the factor prior to model execution, (4) representing the phenomenon explicitly in computer code within the model, and (5) representing the phenomenon implicitly, that is, inclusion only as a logical necessity of some subsuming, larger-scale process present in the model. Among the nine models, the number of explicit or user input representations of human factor process phenomena ranged from 14 to 77. Large-scale models, such as theater-level simulations typically are based upon the results of lower-level models, necessarily implicitly incorporating whatever representations of and assumptions about HPF are contained in the models upon which they are based.

A second issue regarding the representation of HPF in combat simulation models centers upon the extent to which the assumptions made are empirically based. While detailed elucidation by experimentation is generally impossible, other sources of information exist, each with unique shortcomings. For example, Dupuy (1985) and others have proposed greater use of historical data. Of course, any historical account is to some extent both subjective and highly dependent upon specific circumstances. Another source of information is the judgments of experts in the domain in question. Such persons may include combat veterans and behavioral scientists. Problems are that experts frequently disagree and even achievement of consensus does not guarantee accuracy. Another source of relevant information is the data collected from the battalion-size training exercises conducted at the National Training Center (NTC) in recent years (Fobes, 1984). This data, however, is pertinent to only a subset of the HPF of interest and the analysis of data does present some challenges (Whitmarsh, 1985).

A final issue concerning HPF representation is establishing their relative importance in determining model outcomes. Perhaps it is too obvious to state that the HPF which most impact outcomes demand greater attention in terms of both the accuracy and completeness of their representation in combat models.

III. Representation of Target Selection Behavior

The Training Performance Analysis Tool (TPAT) was developed as a part of a larger research program (Banks, 1985) to assist in the analysis and interpretation of data contained in the NTC database. As training resources are always constrained, they require allocation on the basis of expected return. The current work examines how HPF are modelled in TPAT with the purpose of identifying where training intervention is most likely to have the greatest impact. Typically, computer-based combat simulation models are of sufficient inherent complexity that determination of effects of variations in a particular component cannot be assessed by examination . Instead, the model, with changes incorporated, must be executed and the results evaluated. The scope of the current work is limited to the HPF area of target selection, which is relatively richly represented in TPAT.

The design philosophy of TPAT (Weaver and Griesemer, In press) was based upon three principles. First, parameters in the program are primarily based on empirical results obtained from real-time casualty assessment (RTCA) experiments (Clark et al, 1974) conducted by the Combat Developments Experimentation Command (CDEC). Second, the level of detail in the simulation is restricted to only what is necessary to accurately simulate outcomes and the outcomes are limited to those both observable and recorded in the NTC database. Third, Monte Carlo simulation techniques are employed to enable the study of outcome variability. TPAT is simple enough to facilitate understanding and manipulation of HPF without undue time and expense. Further, it may be possible to eventually validate the findings of this investigation using the NTC database.

As implemented, TPAT incorporates three behavioral tendencies of defending forces--friendly forces in the current scenario--which both conflict with Army doctrine and run counter to common sense. They are, however, empirically based (Clark et al, 1974). First, defending tanks and antitank tube-launched optically-tracked wire-guided missiles (TOWS) lack target preferences concerning target vehicle types. Doctrine states that higher threat targets should be engaged first. In the scenario this implies that the long-range opposing force antitank weapons (SAGGERS) and tanks (KTANKS) should be preferred to unarmed armored personnel carriers (APCS) as targets. Defenders presumed interest in self-preservation would also suggest that the APCS, at least at moderate to long ranges, should be targets of secondary importance because they are of relatively little threat.

A second empirically-based behavioral tendency concerning target selection incorporated into TPAT is that newly visible targets are preferred targets. Since new targets would, in general, be less likely to have acquired a target, be more distant, and, therefore, be less threatening than previously visible targets, the same doctrinal principle applies and again is violated. The behavior also appears not to be in the best interests of self-survival.

Finally, TPAT includes the behavioral tendency of defending tanks to prefer previously engaged targets. This preference leads to previously killed targets being frequently engaged. This target perseveration also violates the doctrinal principle of firing at highest-threat targets first. Also, as with the other behavior tendencies, it promotes neither the survival interests of the individual players nor of the defenders collectively.

These three target selection phenomena are included in TPAT because they describe actual behavior of defenders in RTCA experiments. Whether the phenomena exist in NTC training exercises must await the collection and analysis of appropriate data. Rather than speculate about possible psychological bases of these behaviors (e.g. information overload (Miller, 1956) , preference for novelty, need for closure, "shooting gallery mentality," and so forth), it is more important to understand that they are empirically based and to realize that each could potentially be modified to some degree through training. TPAT itself can be used to provide estimates of the effects alterations in these three target selection behaviors have upon simulated battle outcomes. The results should reveal their relative importance on outcomes and, therefore, suggest where the largest training payoff lies.

IV. Modification of Modelled Target Selection Behaviors

TPAT was run employing a digitized portion of NTC terrain with an approximate battalion-sized force (18 tanks and 18 TOWS) in defense and an approximate regiment-size attacking force (30 KTANKS, 30 SAGGERS, and 30 APCS) using rapid approach tactics. To assess the effects of the three behavioral assumptions upon model output, each of the behaviors was modified. The preference for newly visible targets and target perseveration where both reduced by 100 percent. Preference for target type was manipulated by reducing desirability of APCS to zero when there are other target candidates and when the range of APCS was greater than 500 meters. Target selection assumptions were modified for defending tanks and TOWS.

The effects of modifying each of the three target selection tendencies of the defending force are summarized in Table 1. The model scenario was executed fifty times for each target selection modification. Casualty ratios are shown for both the offensive and defensive forces broken down into inflicted and sustained components. The values for casualties can be interpreted as the mean number of vehicle kills per vehicle per trial. The casualties sustained values can be interpreted as the mean probability across trials and across vehicle types of a vehicle being killed.

The results in Table 1 indicate that eliminating either the perseveration on last target or the preference for newly-visible targets has virtually no effect on casualty values in comparison to outcomes from the original model. However, the conclusion is <u>not</u> warranted based upon the the evidence at hand that they have no effect upon performance outcome. Using other scenarios or alternative terrain may reveal they too affect performance. At present, it is appropriate to conclude only that these behaviors have no impact within the

Table 1.

Effects of Changes in Target Selection Processes Upon Combat Outcomes [1]

MEASURE OF PERFORMANCE	ORIGINAL MODEL	NO PERSEVERATION ON LAST TARGET	NO PREFERENCE FOR NEWLY-VISIBLE TARGET	REDUCED PREFERENCE OF APCs AS TARGETS
Offensive Casualty Ratio (inflicted/sustained)	.20/.97	.20/.97	.20/.99	.14/.98
Defensive Casualty Ratio (inflicted/sustained)	2.43/.50	2.48/.50	2.48/.49	2.45/.34

[1] The standard error of estimate across trials is less than .022 for all numbers shown.

current scenario context. By contrast, reducing the preference for APCS as targets produces strikingly different performance. Because defenders concentrate firepower on KTANKS and SAGGERS, these high-threat attackers are attrited at a relatively higher rate. This leads to a reduction in the casualties the attackers are able to inflict. In turn, this produces the result that the likelihood of defender being killed falls from .50 to .34. The values represent a decrease in per vehicle risk of approximately a third.

The finding that reducing the attractiveness for APCS as targets produces fewer defender losses provides unambiguous support for the doctrinal principle that targets of greater threat should be engaged first. Also of interest is the supposition that target selection behavior is probably amenable to modification through appropriately designed and conducted training. While target discriminability probably has inherent limits, especially at extreme range, it is plausible to expect some improvement with training.

A series of TPAT runs was conducted to explore how much effect upon outcomes might be expected from varying degrees of reduction in the preference for unarmed APCS as targets. In addition to the preference level of APCS in the original model and the previous 100 percent reduction, their attractiveness was reduced by 25, 50 and 75 percent. Fifty runs under each condition were performed. Figure 1 presents the results displayed as mean probabilities of defensive vehicles becoming casualties. The specific values are less meaningful than the overall pattern of the relationship. Finding a nonlinear relationship underscores the need for actually making model changes and executing the model to determine effects. In addition, the direction of the nonlinearity observed suggests that small amounts of reduction in the preference of APCS as targets yield little reduction in risk--a 25 percent reduction from the original model produces no risk reduction. The relationship suggests that only when there is 50 percent or greater reduction in preference are substantial benefits realized. This finding suggests that while relevant training may reduce risks to defenders, incomplete or ineffective training may produce no benefit.

V. Conclusions

Clearly, assumptions made about HPF can affect the outcome of simulated combat. This finding has training as well as analysis implications. Investigation of HPF effects in combat models can be important to the identification of training needs, to the planning of training, and to the development of training criteria goals. Model developers may be aware of the

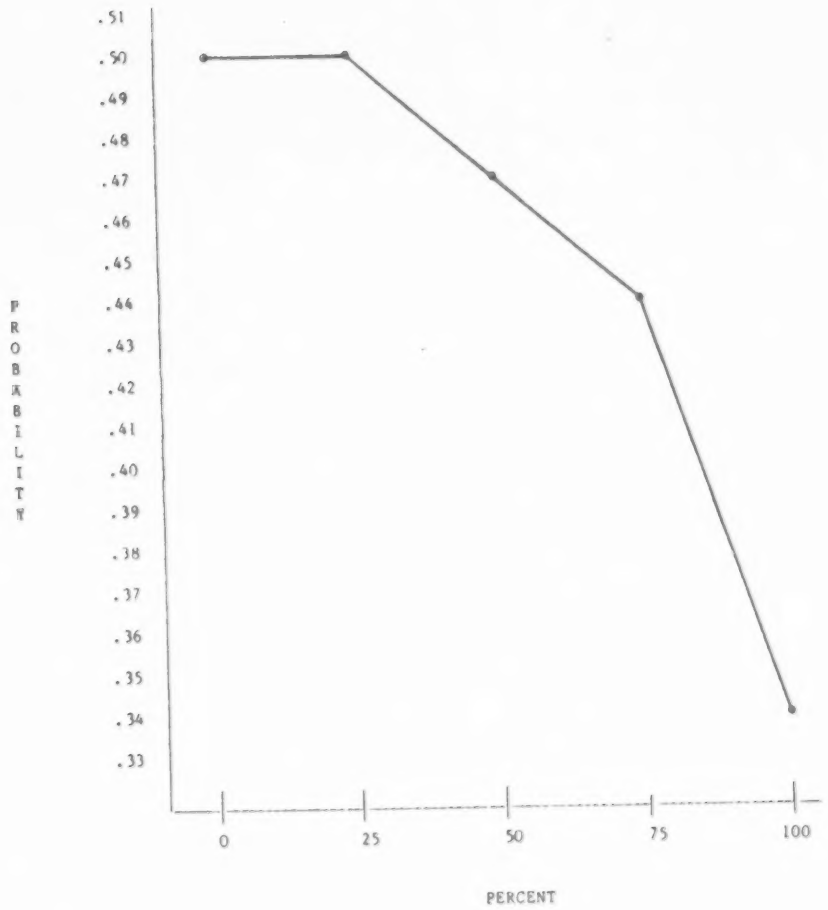


Figure 1. Effect of percent reduction in defender target preference for APCs upon mean probability of a defending vehicle becoming a casualty.

HPF assumptions and simplifications that have been made and the consequential limitations upon model outcomes. It is also vitally important for policy decision makers, the model users, to know of these assumptions and caveats. Therefore, the value of models is enhanced when these considerations are explicitly documented.

Further systematic research on the HPF representation in combat simulation models is necessary. Several of the specific needs are: (1) development of strategies for gathering relevant empirical data, (2) determination of most important HPF, and (3) identification of methods for better incorporating the effects of HPF into combat simulation model findings.

VI. References

Banks, J.H. (1985). An overview of ARI's research program on the National Training Center. In the Proceedings of the Military Testing Association, San Diego, CA.

Brewer, G.D. & Shubik, M. (1979). The War Game: A Critique of Military Problem Solving. Cambridge, MA: Harvard University Press.

Clark, D.B., Marchi, R.P., Sargent, J.D. Vanarsdall, D. & Weaver, W.B. (1974). Final Report, TETAM Extended Analysis, Vol. I. BDM/CARAF FR-74-075.

Dupuy, T.N. (1985). Criticisms of combat models cite unreliable results. Army, 35(3), 16-18.

Fobes, J.L. (1984). National Training Center Data Handbook. ARI Research Product 84-17, Alexandria, VA: U.S. Army Research Institute.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Miller, G.J. & Bonder, S. (1982). Human Factors Representations for Combat Models. ARI Technical Report 571, Alexandria, VA: U.S. Army Research Institute.

Weaver, W.B. & Griesemer, H.A. (in press). NTC Training Performance Analysis Tool. Research Note, Alexandria, VA: U.S. Army Research Institute.

Whitmarsh, P.J. (1985). Types and quality of National Training Center data. In the Proceedings of the Military Testing Association, San Diego, CA.

QUALITY OF CANADIAN FORCES OTHER-RANK RECRUITS: TRENDS 1975-1983

Major K.W.J. Wenek

Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario, Canada

Introduction

Depending on whether a country chooses to man its armed forces on the basis of voluntary or obligatory service, the strategic forecasting of human-resource supply and the planning of human-resource needs will require different emphases. Conscript forces, which are more or less ensured a reasonable supply of recruits and a representative share of the available youth talent, are primarily concerned with personnel screening and allocation functions. All-volunteer forces (AVFs), on the other hand, which are particularly susceptible to the effects of free-market economics on the quantity and quality of the available manpower pool, must devote considerable attention to the business of attracting applicants before practical screening and allocation policies can be implemented. Moreover, because AVFs are dependent on the willingness of young people to serve in the military, special attention must be given to any qualitative biases in the composition of the target recruiting pool and recruit intake introduced by the market forces which directly affect enlistment propensity. (For a more comprehensive discussion of these issues, see Cotton, Crook & Pinch, 1978; and Pinch, 1982.)

While there are a number of dimensions on which qualitative comparisons among recruit cohorts can be made, military staffs are typically concerned with the trainability of recruits, that is, their potential to acquire those job skills which are a necessary, but not a sufficient, condition for operational effectiveness. In the Canadian Forces (CF), the general indices of quality and trainability most commonly invoked are entry scores on the General Classification (GC) test (a group-administered intelligence test) and educational attainment. Although the relationship between these indices is not entirely orthogonal, GC test scores can be essentially interpreted as measures of learning potential, while educational attainment levels provide measures of acquired learning. Trends in the quality of Other Rank (OR) recruits over the 1975-83 period will be examined in these terms.

Trends in OR Recruit Quality: 1975-1983

As illustrated in Table 1, the average GC-test scores of OR recruits and the distribution of these scores have been remarkably consistent over the 1975-1983 period. Furthermore, because the GC test is a standardized test, it seems reasonable to conclude that recruits

---

The views and opinions expressed in this paper are those of the author and not necessarily those of the Department of National Defence.

enrolled in 1983 were, in an absolute sense, neither more nor less able or trainable than those enrolled in previous years.

Table 1

Raw GC Score Means and Standard Deviations
for Anglo and Franco OR Enrollees (1975-83)

Intake Year	Anglo Enrollees		Franco Enrollees	
	Mean	SD	Mean	SD
1975	42	9.8	37	8.9
1976	44	9.5	40	8.6
1977	44	9.3	40	8.6
1978	45	9.4	40	8.3
1979	44	9.2	40	8.4
1980	44	9.4	39	8.3
1981	43	9.3	38	8.4
1982	43	9.4	39	8.5
1983	44	9.2	40	8.3

A different pattern emerges when educational attainment data are considered. As shown in Table 2 (Tivendell & Gaudet, 1985), the distribution of educational attainment levels among CF enrollees over the 1975-83 period has shifted towards the better-quality end of the spectrum, with higher proportions of enrollees possessing some, or complete, post-secondary and university education and lower proportions of enrollees

Table 2

Percentages of CF Enrollees at Various Education Levels
Compared to Population Percentages

Year	Education Level							
	Some/Complete Elementary		Some/Complete Secondary		Some/Complete Post Secondary		Some/Complete University	
	CF	Cdn Pop	CF	Cdn Pop	CF	Cdn Pop	CF	Cdn Pop
1975	7.0%	5.5%	85.9%	60.3%	3.3%	25.2%	3.7%	3.5%
1976	4.9%		87.4%		4.2%		3.5%	
1977	5.0%		86.9%		4.3%		3.9%	
1978	3.7%		87.7%		4.1%		4.5%	
1979	3.3%	4.9%	87.7%	67.2%	5.0%	19.7%	4.1%	3.4%
1980	4.1%		87.8%		4.7%		3.5%	
1981	4.3%		88.1%		4.4%		3.2%	
1982	3.3%		83.1%		7.7%		5.8%	
1983	1.9%	4.1%	81.7%	64.3%	9.8%	25.2%	6.4%	3.1%

with some, or complete, elementary and secondary education evident. In comparison to benchmark figures for the Canadian population, CF enrolments from the elementary-education group have shifted over time from over-representation to the more favourable position of under-representation. Similarly, in the university-education group, the CF has maintained a very favourable recruiting position, and has, in fact, improved the extent of over-representation. It is in the middle groups, however, which are the sources of the bulk of its recruits, that the CF has had less success, continuing to be over-represented at the secondary education level and substantially under-represented at the post-secondary level.

Data obtained from sources (Tierney, 1981; Park, 1982) which deal exclusively with OR educational attainment, and which break out educational attainment figures in terms of program completion and non-completion, provide a more useful, if less complete, picture. While Table 3 confirms the general shift towards higher proportions of better educated recruits and lower proportions of the less educated, it also reveals certain weaknesses in the CF's recruiting posture. Recruit intakes continue to be over-represented by non-high school graduates and under-represented by graduates of post-secondary schools and universities. These trends are reason for concern in view of the training demands imposed by the proliferation of complex technology in all areas of military functioning and in light of the strong relationship between high school completion and early-career survival (Sinaiko & Schefieu, 1982).

Table 3

Percentages of CF OR Enrollees at Various Education Levels Compared to Population Percentages

Year	Education Level					
	Non-High School Grad		High School Grad		Post Sec & Univ	
	CF	Cdn Pop	CF	Cdn Pop	CF	Cdn Pop
1975	81%		17%		2%	
1976		44%		20%		36%
1980	60%		29%		11%	
1981		42%		21%		37%
1982	48%		----------52%----------			
1986*		39%		19%		42%

* Projected

Discussion

The information on trends in educational attainment presented above prompt at least two major questions. Why has the CF been relatively unsuccessful in attracting a representative share of technical/community college students? Do the apparent general gains in the educational quality of recruits represent real and sustainable gains?

Surveys of applicants to the CF (Tierney, 1981) and of high school students planning to pursue post-secondary education (Urban Dimensions Group, 1982) have indicated that interest in the CF as a career option generally decreases with higher levels of education and that, among technically oriented students in particular, the CF is relatively unattractive as a training or employment opportunity. For those high school students who have the potential and desire to acquire technical skills fairly rapidly, the CF has very little to offer in the way of accelerated training programs or incentive programs, and it is in this sense that the CF has been displaced as a major technical training institution by the numerous technical institutes and community colleges which sprang into being during the 1960s and 1970s (Pinch, 1982; McMenemy, 1984). With respect to those who already hold post-secondary qualifications, the CF does not have a lateral-entry program for skilled applicants nor does it yet have a career structure which specifically rewards technical-skill specialization (although a program to address this latter area is under development). Overall, therefore, it must be concluded that, unless the CF's technical training programs, career structures, and benefit packages can be made more attractive to aspiring or qualified technicians and technologists, and can be made more competitive in comparison to civilian educational and career opportunities, this important source of personnel is not likely to be adequately tapped for some time to come.

The question of apparent versus real gains in educational quality among OR recruits is somewhat more difficult to answer directly. As a point of departure though, it should be noted that the pass/fail criteria used in schools and which, to a large extent, affect the rate and extent of school advancement, can vary considerably over time, particularly when institutional survival is threatened by declining enrolments. This elasticity of standards and the resulting difficulties in making absolute comparisons in achievement should also be viewed in the light of recent public controversy in the United States and Canada over the declining quality of education. Research on trends in scholastic achievement scores (standardized, hence comparable, measures) strongly indicates a decline in the academic competence of youth (Waters & Laurence, 1982). What these observations collectively suggest is that neither the absolute nor relative quality of education in recent years is what it used to be.

With respect to the sustainability of, or improvement upon, the present state of recruit educational quality, future developments will depend on a number of factors, many of which have contributed to the past few years of good fortune. Since the onset of a national economic recession in 1981, the national unemployment rates in the work force have ranged between 10% and 13%, with the rates running somewhat higher, at 16% to 20%, for the youth population (Statistics Canada, 1985). Consequently, the effects of the bleak demographic prophecies of the late 1970s have been temporarily postponed. Over the same time period, labour-market conditions have had a substantial impact on the voluntary attrition of CF Other Ranks. As reported by Mendes and Lyon (1984), attrition fell off rapidly from 1979 to 1983 as unemployment rose ($r=-.74$), and annual recruiting quotas fell accordingly. The combined effects of an increasing

supply, brought about by high unemployment in the 15-24 year-old segment of the population, and of a decreasing demand, brought about by declining attrition, can best be seen in the applicant and enrolment figures for this period and the corresponding dramatic improvement in CF selection ratios (Table 4).

Table 4

Total Number of CF Other Rank Applicants and Enrollees for Fiscal Year 1979/80 to 1983/84

Fiscal Year	Applicants	Enrollees	Selection Ratio
1979/80	28,468	10,316	1:2.8
1980/81	35,248	11,866	1:3.0
1981/82	39,626	12,261	1:3.2
1982/83	40,148	6,674	1:6.0
1983/84	29,738	4,038	1:7.4

It would be extremely surprising if, under these conditions, recruit quality were anything but high. Yet because current CF recruit quality is largely an artifact of a recent national economic downturn, it is sobering to think what might happen if there is an upswing in the economy. On the supply side, declining unemployment in the youth population and the intractable effects of demographics will rapidly shrink the eligible personnel pool. On the demand side, a surge in voluntary attrition, as external employment opportunities materialize, will probably drive up recruiting quotas to or beyond pre-recession levels. The first impact would be on quality. Numbers could also be a problem.

Without forward-looking recruiting strategies and personnel policies, particularly those designed and preplaced with a view to coping with a manpower-shortage contingency, the CF runs a real risk of suffering a strategic self-inflicted wound. The policy areas which require urgent review in order to forestall this eventuality include the following:

a. redefining the eligible manpower pool, which would minimally entail the removal of age barriers, the expansion of women's roles, and the development of lateral-entry programs for skilled applicants;

b. improving the fit between military training and civilian education, which could mean increased reliance on "off-the-shelf" skills and the establishment of CF-sponsored programs at technical colleges;

c. developing the recruitment potential of the military-socialization infrastructure, which basically means providing more support to cadet and Reserve organizations; and, finally,

d. exploring the possibility of a national service scheme, which could vary in application from the institution of a national public-service draft, which includes military service as an option, to a comprehensive incentive-and-benefits program for military service.

REFERENCES

Cotton, C.A.U., Crook, R.K., & Pinch, F.C. (1978). Canada's professional military: the limits of civilianization. Armed Forces and Society, 4, 365-90.

McMenemy, J.P. (1984, April). Recruiting, selecting, and training the next generation of technicians. Paper presented at the NATO DRG Panel VIII Workshop on Applications of Systems Ergonomics to Weapons Systems Development, Shrivenham, England.

Mendes, H.C., & Lyon, C.D.F. (1984). Canadian Forces attrition/retention study (Working Paper 84-8). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Park, R.E. (1982). Trends in recruit quality: An analysis of enrollee profiles across recruit zones from July 1980 to June 1982 (Technical Note 5/82). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Pinch, F.C. (1982). Military manpower and social change: assessing the institutional fit. Armed Forces and Society, 8, 575-600.

Sinaiko, H.W., & Scheflen, K.C. (1982). Correlates of first term attrition: A comparison across TTCP nations. TTCP Technical Panel UTP-3.

Statistics Canada (1985). The labour force: July 1985. Ottawa: Supply and Services Canada.

Tierney, E.C. (1981). The applicant survey 1979/80: A preliminary analysis of national and regional effects (Working Paper 81-4). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Tivendell, J., & Gaudet, J.W. (1985). Socio-demographic trends and related changes in Canadian society affecting the Canadian Forces personnel supply (Research Report 85-2). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Urban Dimensions Group Inc. (1982). Occupational and educational plans of Canadian youth as they impact on ROTP subsidized education opportunities (Research Report 82-9). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Waters, B.K., & Laurence, J.H. (1982). A comparison of test score trends: Civilian versus military examinees and recruits (1972-1981). Offprint.

Role, Importance and Availability of GFAF Reservists

Heinz-J. Ebenrett
Federal Armed Forces Office, Bonn, FedRep Germany
presented by
Hans Kuessner
Federal Armed Forces Office

The Mission

The mission assigned to the German Federal Armed Forces (GFAF) by NATO requires the three Services - Army, Air Force, and Navy - to render an essential and decisive contribution to the collective defense of Central Europe and the Baltic Approaches.

With its 345,000 soldiers on active duty, the Army contributes the largest share to the total strength. The Air Force will grow from almost 111,000 active airmen in peace to about twice that size in war while the Navy will increase its present complement of 39,000 sailors 1.75 times. These strength figures can only be achieved by armed forces that are structured as a conscript army with a high percentage of temporary-career and regular personnel and that are supported by strong reserve personnel.

Reserve Personnel

Since the inception of the GFAF, 5.2 million men have performed military service in the "Bundeswehr". 2.4 million of them are still available for military purposes due to age and level of training.
At present, 762,000 reservists are included in plans for assignments which must be filled by mobilization to achieve the full defensive capability of the armed forces. Of these 762,000 reservists firmly planned for mobilization, 40,000 are officers, 196,000 NCOs and 526,000 privates.
In the next two years, it is intended to raise these numbers by 90,000 in order to be able to perform the task of the Wartime Host Nation Support (WHNS) Program. As a result, wartime strength will grow to 1.34 million service personnel, beginning in 1987.

Use of Reserve Personnel

Army and Air Force are currently providing for 60,000 reservists in the "Standby Readiness" component. In a crisis, the Federal Minister of Defense can recall standby readiness personnel to active duty in order to improve the operational readiness of specific combat units of the Army and the Air force. Following a decision by the Federal Government, the men in the Alert Reserve are recalled in the course of mobilization measures. These men will round out units and agencies of all the Services, assume tasks designed to maintain the operational freedom for the armed forces, and support the allied forces. The balance of the

reservists firmly included in the plans will be available to major units as replacement personnel.
The requirement of the armed forces for reserve personnel can be met in terms of numbers, but there are still some deficiencies as far as quality is concerned, since it has so far been impossible to train all reservists during their periods of active military service for their particular wartime assignments. They must be given additional training during reserve duty training periods.

Reserve Duty Training

Training and extension training of reserve personnel is provided by a system of reserve training periods that is geared to the needs of wartime operational readiness:

Individual reserve duty training periods are designed to provide extension training to the individual reservist. Their normal duration is two to four weeks. Reservists often volunteer for them.

Mobilization exercise of a duration up to 12 days are in the first place intended for the training of those elements which in peacetime exist merely as equipment holding units with no personnel or as cadre-strength units.

Mobilization alert exercises are conducted without warning. They last up to three days, and their purpose is to practice mobilization procedures.

The importance which the armed forces attach to the reserve component will continue to grow in the years to come. The reservist concept is currently being updated in connection with studies under way concerning the structure of the German forces. The goal is not only to continue to meet the numerical requirement of the forces for reserve personnel, but also to improve the quality of training through organizational measures and to ensure that the burdens involved are distributed to all reservists as equitably as possible. Spaces for reserve duty training are to rise to 6,600 by 1986. This will make it possible, for the first time to recall more than 200,000 reservists per annum for reserve duty training periods and to narrow the still existing gaps in their training.

In the nineties, the number of spaces for reserve duty training will gradually be increased to 15,000. These measure will enable up to 400,000 reservists per year to be recalled for reserve duty training periods and to be trained for their wartime assignments. Reserve personnel must then expect to be recalled more frequently and perhaps for shorter periods of time. Such a conscripted program would pose an additional burden both on the individuals concerned and on industry and economy.

Model "Reservist Volunteering by Particular Declaration"

There are doubts that these far reaching plans (mentioned above) can be realized; especially with regard to the readiness of reservists to participate repeatedly on extension training

periods.

In order to attract qualified reserve personnel for matters of intensified training, a special model has been developed. This model is called "Reservist Volunteering by Particular Declaration". It consists of the following elements:

1. The reservist formally declares to render repeatedly (additional to obligatory reserve duty training) mobilization exercise and training periods of at least 28 days annually (maybe less when realized) for a minimum of 3 years.

2. The status of the "Volunteering Reservist" will be that of a conscript and his payment shall be fixed according to his civilian income.

3. With respect to dates and terms of exercise the reservist is offered to come to an understanding with his mobilization unit.

4. As for incentives the reservist may expect development and promotion in his individual military career.

The model primarily aims at leadership personnel in mobilization units up to the level of a battalion commander. The "Volunteering Reservist" will generally receive military training in the same mobilization unit. The purpose of this is to enhance unit cohesion among reservists as a way of increasing unit effectiveness. Additionally it is expected that the commitment of the "Volunteering Reservist", i.e. their understanding of and corresponding to common defense duty, will have a distinctively positive influence on fellow citizens.

(Supplementary to the model mentioned above there are reflections about the introduction of a status for part-time soldiers. Preliminary considerations aim at civilian operators, mechanics, technicians etc., who can make it possible to serve as part-time soldiers; for example one day per week or several days a month. That type of "reservist" would be earmarked to move and maintain military equipment and material of those elements, which in peacetime exist merely as equipment holding units with no personnel or as cadre-strength units. Nevertheless, at the present time there is no legal basis for part-time soldiers in Germany and therefore this model has not yet exceeded the status of preliminary considerations. If at all, it may be carried on when the model concerning the "Volunteering Reservists" has been realized.)

Inquiry of the degree of Acceptance

The decision whether the "Volunteering Reservist" model will be realized or not is still is pending. Inter alia, it depends on the degree of acceptance among the reservists. In order to get valid data for matters of prognosis and planning, the

Psychological Service of the GFAF has been assigned the mission of determinig this acceptance by means of a representative survey.

Currently a representative sample of 2.500 reservists is being conducted. Results will be given not before the end of the year. Nevertheless, some preliminary trends can be drawn from pretest data. The following preliminary data may be of some interest.

Primarily for matters of testing the suitablility of the questioning instrument a sample of 182 reservists in leadership functions, who happened to render a reserve duty training in August 1985 had been asked about their interest in the "Volunteering Reservists" model. The proportion of answers to the central question show a distinctly high degree of acceptance:

Interest in Participating:(N)

	total	NCO's or cand.	Senior NCO's	Officers
"yes"	74	33	16	25
"uncertain"	22			
"no"	85	63	30	15
	(1 missing)			

74 of the 182 reservists (= 41 %) stated an individual interest, the officers in the sample even by majority.

Reason for Rejection

With respect to those subgroups who showed "no" or "uncertain" interest (N = 85 + 22 = 107) it is obvious that the predominant reasons given for rejection referred to occupational claims or hinderances. Claims of family were of minor but significant importance. Although each third of the uninterested reservists manifested doubts in the purpose and sense of military duty, too, it can be said that the main reasons for rejection are more objective hinderances in the occupational and personal sphere and less negative attitudes or reservations towards the military duty:

Reasons for Rejection(multiple answers)
Subgroups: "no interest" (N=85) and "uncertain" (N=22)

	"primarily	"important	(others)
occupational claims	63	18	26
reservations of the boss	28	23	53
occupational drawbacks	30	15	62
claims of family	31	27	49
reservations of the spouse	21	26	60
doubts in purpose and sense of military duty	20	16	71

Reasons for Volunteering

As for the subgroup of those reservists who showed interest in the model it is noteworthy that the actual source of that interest seems to be a positive attitude towards the military in general. The far overwhelming majority of the interested reservists stated a noticeable rate of contacts to the armed forces, positive experiences during terms as well as the utility of military training while on civilian business, The reasons which may attract them to consider a participation on the "Volunteering Reservist" model are given in the following:

Reasons/Conditions for Volunteering
Subgroup: "interested in . . . " N = 74

	"very important"	"important"	(others)
agreements upon dates	53	18	3
good morale in the unit	36	31	7
suitable duties	36	23	15
military leadership development	34	25	15
promotion in rank	20	27	27
actions near home	25	14	35
financial incentives	15	30	29

The condition to come to an understanding with the mobilization unit upon the dates and terms of training periods is of utmost importance. Nearly all of the interested reservists wanted to be able to plan and determinate duty terms in advance. If that is made sure, they seem to be more attracted by idealistic views and incentives (i.e. morale, duty, leadership) and less by objective advantages (i.e. promotion, payment, short distances).

Summary

Although the data may not be considered representative, they arre suggestive of certain tendencies, which can specifically be verified in the main study. With respect to the far-reaching aims of the "Volunteering Reservist" model it is hoped that both substantial trends we mentioned above may be confirmed; i.e. the sufficiently high degree of acceptance as well as the note that in the first line reservists with idealistic views and motives are attracted. Therefore, their commitment and understanding of duty to render repeatedly voluntary training periods may justly be expected to have a lasting positive influence on defense motivation of draftees and civilians, alike, in their social surroundings.

Retirement Readiness as a Function of Transition Assistance and Trade

Major F.P. Wilson
Canadian Forces: Director of Personnel Selection, Research and Second Careers

Retirement normally represents the end of routine daily employment in order to earn one's living. Mid-career change, on the other hand, involves leaving one job, either voluntarily or involuntarily, with the expectation of undertaking other employment. Although leaving the military after serving 20 or more years is commonly referred to as retirement, the majority of Canadian Forces (CF) leavers go on to second careers (Pinch & Hamel, 1978). Better insight into the problems of military mid-career change can be acquired by considering this phenomenon from the perspective of developmental and vocational psychology theories.

Research completed by Levinson, Darrow, Kline, Levinson and McKee (1978) suggested that there is a human life cycle made up of a series of stages and substages, with everyone more or less passing through this process. Levinson et al. (1978) and others (Lowenthal, Thuristes & Chiriboga, 1975; Neugarten, 1977; Vaillant, 1977) view the 40 to 50 age period as being particularly volatile, frequently marked by personal crises. Career and occupational concerns play a significant role in theories of adult development. Personal problems emanating from the workplace can gain overriding importance in a person's life. Medical research in the US Forces identified the "retirement syndrome" (Berkey & Stoebner, 1968; Druss, 1965; Greenburg, 1965; McNeil & Griffen, 1967; Milowe, 1964) which is characterized by both social and intrapsychic problems. These writers have posited that the social difficulties stem from: the financial inadequacy of the annuity, which must be supplemented by other income from a second career in the civilian labour force; indecision as to where to settle upon leaving the forces; loss of friends with similar jobs and common interests; and, perceived difficulty in finding employment in an environment that harbours misguided stereotypes of what the military retiree has to offer. The intrapsychic problems are in part attributable to the anticipated loss of status and high degree of responsibility for men and equipment, loss of the security which guarantees the individual that he and his family will be cared for if he becomes ill or incapacitated, loss of friends and lifestyle structure, and the spectre of impending ambiguity, uncertainty, complexity, and conflict attached to "starting over" in a civilian career.

According to Druss (1965) and Greenburg (1965), these social and intrapsychic problems during the latter part of the serviceman's career lead to marital discord, excessive drinking, depression, insomnia, a variety of psychosomatic complaints, and reduced work performance. Researchers in the Canadian context have also examined problems associated with mid-career change. Pinch and Hamel (1978) pointed out that many long-term service-members lack knowledge concerning the civilian job market, are unable to meet formal educational prerequisites for jobs, and have failed to recognize the necessity for advanced pre-retirement planning. These authors also found that personnel from "hard" military occupations (i.e. trades without readily discernable civilian counterparts) were unemployed longer and were

less likely to regain former pay and status levels. Second career concerns affect all members to some degree and can have a deleterious effect on personnel performance during the last several years of service.

Recognizing that providing pre-retirement assistance not only fulfills a moral obligation, but also impacts favourably on operational effectiveness, the Department of National Defence (DND) has instituted the Second Career Assistance Network (SCAN) program. Components of SCAN include instruction in resume preparation and job search techniques, information on financial planning, military/civilian trade accreditation, and, counselling on personal and social adjustment concerns judged to be significant to mid-life career change. The aim of the program is to help the participants become more psychologically and practically prepared to leave the military and begin a second career. Even though SCAN has been functioning since 1978, its effectiveness for readying personnel for retirement has yet to be established. The work reported in this paper is part of research being conducted to assess SCAN's effectiveness in aiding the second career transition process.

To examine the effects of SCAN, the psychological construct vocational or career maturity was considered to be an appropriate conceptual framework. In previous research, career maturity was shown to be a major determinant of educational and occupational success for adolescents (Super, Crites, Hammel, More, Overstreet & Warrath, 1957; Super & Overstreet, 1960; Zelkowitz, 1974). Recognizing the difficulty inherent in measuring vocational maturity beyond adolescence, Super (1977) presented a revised theoretical model better suited to understanding career development in adults. Later modifications (Super & Knasel, 1979; Super & Kidd, 1979; Super, 1983) operationalized the model's dimensions and recommended that, in the case of adults, the term "career adaptability" be used instead of vocational maturity.

Thus, according to Super's adult paradigm, career adaptability is made up of five dimensions: Planfulness, Exploration, Information, Vocational Decision-Making, and Reality Orientation. This paper will concentrate on the Exploration or Exploratory Behaviour (EB) dimension of the model. EB is a broad concept concerned with attitudes toward vocational resources, ability to distinguish their worth from a personal perspective, and willingness to utilize them in searching for a second career. Due to the unique context of this research, the more descriptive term <u>military disengagement readiness</u> (MDR) is used synonomously with career adaptability throughout the paper.

Inasmuch as career adaptability or military disengagement readiness is a multi-dimensional psychological concept (Jordaan & Heyde, 1979; Super & Overstreet, 1960; Super & Knasel, 1979), it provides an appropriate medium through which the effects of the SCAN program can be measured. If SCAN is effective, there should be a positive relationship between participation in the program and MDR. For those with marketable civilian skills (e.g., electrician, plumber), SCAN appears to offer services that are immediately useful (e.g., resume writing, trade certification, job search techniques). However, for those whose trades are only remotely marketable, such as Combat Arms, the services offered by SCAN (e.g., information on academic upgrading,

skill retraining courses) offer a more long term solution to their problems. This study treated the SCAN program and military trade category as independent variables and measured their effects upon one of the disengagement readiness dimensions, i.e., the independent variable, EB. More specifically, this research investigated: a) whether there are different initial levels of MDR, as measured by the EB scale, across military occupations; b) the overall effect of SCAN on EB; and, c) whether trades were differentially effected by SCAN.

METHOD

Subjects were 354 male non-commissioned officers divided into two groups: the SCAN treatment group, those who were registered in the program and had attended at least one seminar; and, the No SCAN control group, those who had no contact with the program but intended to register in SCAN before retiring. Within each of these groups, three trade categories were imposed: Cat 1, "hard" military trades with no civilian counterpart; Cat 2, low to semi-skilled trades with civilian counterpart; Cat 3, highly skilled trades with civilian counterpart. The three trade categories were developed from occupational descriptions contained in the Canadian Classification Dictionary of Occupations (CCDO). Cat 1 contained 18 trades, Cat 2 contained 39, and Cat 3 contained 55.

The Military Disengagement Readiness Inventory (MDRI) was designed to measure disengagement readiness within CF retirees. The MDRI questionnaire contained eight scales reflecting Super's dimensions of career adaptability. The EB scale was developed based on the Super and Knasel (1979) and Jordan (1963) operational definition of vocational exploratory behaviour. It sought to measure the individual's involvement in such second career planning areas as learning about job search techniques, writing vocational tests, determining pension benefits and obtaining information about retraining programs. The scale contained 16 items and was scored using successive categories from 1 to 6. An item analysis indicated a coefficient alpha of .9038 for this scale.

The SCAN group was administered the MDRI at the start of a SCAN seminar, and again two months afterwards. To generate a control group, the MDRI was sent out during the same timeframe to members identified as having five or less years to serve before retirement. Individuals who indicated on the first MDRI that they were planning to register in SCAN at a later date were selected as the No SCAN group and received a second questionnaire at the same time as the SCAN group.

ANALYSIS AND RESULTS

A split plot ANOVA was used with two between subject factors, i.e., trade category and treatment group. The within subject variable, EB, was measured via the two MDRI administrations approximately two months apart. During the intervening period, the SCAN treatment group attended a SCAN seminar and participated in other program activities, whereas the No SCAN group was not exposed to any aspect of the program.

TABLE 1

Mean Exploratory Behaviour Scores of Treatment Group Over Time*

GROUP	TIME 1	TIME 2
SCAN (N=270)	47.87 SD=17.51	57.04 SD=18.45
No SCAN (N=84)	39.85 SD=16.28	42.45 SD=16.83

* Trade group means are not shown as all differences were non-significant.

Mean before and after EB scores for the SCAN and No SCAN groups are presented in Table 1. Overall F value indicated that there is no significant trade by treatment interaction over time. However, the treatment by time interaction was significant, F (347,1) = 14.60, $p < .001$ with the group receiving SCAN showing a greater increase in EB on the second measure of the MDRI. There were also significant main effects for time, F (347,1) = 108.11 $p < .001$, and treatment, F (347,1) = 30.68 $p < .001$, but not for trade, F (347,2) = .4759, $p > .05$. (Reported F values are based on multivariate test results.)

Paired comparisons, using Dunn's test of significance ($p < .05/4 = .0125$) indicated that upon the first testing, there was a significant difference between the SCAN and No SCAN groups, with the SCAN group displaying greater EB, t (359) = 3.57, $p < .001$. The second administration showed even greater difference between these two groups, t (359) = 6.46, $p < .001$. Over time, the SCAN group showed a significant increase in EB, t (269) = 10.63, $p < .001$, whereas the No SCAN group remained unchanged, t (83) = 1.91, $p = .06$.

DISCUSSION AND CONCLUSION

The aim of this research was to determine whether members trained in trades having readily recognizable civilian counterparts are more prepared to make the transition to a second career, whether SCAN has an impact on the readiness of the individual to make the transition, and, whether SCAN differentially affects readiness to seek a second career across trade categories. As demonstrated in this research, SCAN is definitely affecting an individual's predisposition to explore various ways of finding second careers. This corroborates earlier research conducted with adolescents demonstrating that career maturity can be taught (Yates, Johnson, & Johnson, 1979). However, all trades appear to be affected equally by SCAN. That is, those in "hard" military occupations report the same exploratory behaviour as those in more readily marketable occupations. An explanation for this finding may be that the EB items have more obvious applicability to second career concerns than other MDRI scales. Such activities as preparing a career resumé, learning about job search techniques, and writing vocational tests would appear to have equal relevance across trade categories.

However, this finding also may be an indication that the military occupations could be rearranged on a rational or statistical basis to reflect a better delineation of categories.

The results of this research suggest that SCAN is having a positive affect on at least one aspect of second career adaptability. It seems likely that an increase in knowledge and skills directed at easing the transition to the civilian workforce would reduce the stress and anxiety associated with retiring from the military. Thus, this leads to the belief that not only is SCAN fulfilling a moral obligation, but it likewise may be helping to maintain a more effective servicemember over the last several years prior to retirement.

This paper reported on the exploration dimension of the second career adaptability concept (Super, 1983). Although EB is only one of five discrete dimensions encompassed by second career adaptability, the results of this research suggest that the construct can be operationalized and used for gaining greater understanding of military second career transition. However, much stronger evidence to support the validity of the construct would be provided by investigating the post-retirement status of SCAN and non SCAN retirees.

REFERENCES

Berkey, B.R., & Stroebuer, J.B. (1968). The retirement syndrome: A previously unrevealed variant. Military Medicine, 133, 5-7.

Druss, R.G. (1965). Problems associated with retirement from the military service. Military Medicine, 130, 251-255.

Greenburg, H.R. (1965). Depressive equivalents in the pre-retirement years: The old soldier syndrome. Military Medicine, 130, 251-255.

Jordaan, J.P. (1963). Exploratory behaviour. In D.E. Super, R. Statisherky, N. Martin, & J.P. Jordaan. Career Development: Self-concept theory. New York: College Entrance Examination Board.

Jordaan, J.P., & Heydé, M.D. (1979). Vocational maturity during the high school years. New York: Teachers College Press.

Levinson, D.J., Darrow, C.M., Kline, E.B., Levinson, M.H., & McKee, B. (1978). The seasons of a man's life. New York: Alfred Knopf.

Lowenthal, M.F., Thurister, M., & Chiriboga, D. (1975). Four stages of life: A comparative study of women and men facing transitions. San Francisco: Jossey-Bass.

McNeil, J.S., & Giffen, M.B. (1967). Military retirement: The retirement syndrome. American Journal of Psychiatry, 123, 848-853.

Milowe, I.D. (1964). A study in role diffusion: The chief and the sergeant face retirement. Mental Hygiene, 48, 101-107.

Neugarten, B. (1964). Personality in middle and late life. New York: Atherton Press.

Pinch, F.C. & Hamel, C., (1978). The transition to civilian life among CF members: Preliminary results, stage II. (Report 78-3). Toronto: Canadian Forces Personnel Applied Research Unit.

Vaillant, G.E. Adaption to life. (1977). Boston: Little, Brown and Co.

Super, D.E. (1977). Vocational maturity in mid-career. Vocational Guidance Quarterly, 25, 294-302.

Super, D.E. (1983). Assessment in career guidance: Towards truly developmental counselling. The Personnel and Guidance Journal, May 1983, 555-562.

Super, D.E., Crites, J.O., Hummel, R.C., Moser, H.P., Overstreet, P.L., & Warnath, C.F. (1957). Vocational development: A framework for research. New York: Teachers College Press.

Super, D.E., & Overstreet, P.L. (1960). Vocational maturity of ninth grade boys. New York: Teachers College Press.

Super, D.E., & Kidd, J.M. (1979). Vocational maturity in adulthood - Toward turning a model into a measure. Journal of Vocational Behaviour, 14, 255-270.

Super, D.E., & Knasel, E.G. (1979). The development of a model, specifications, and, sample items for blue collar workers. Final report to Canada Employment and Immigration Commission. National Institute for Careers, Education and Counselling, 1979.

Yates, C., Johnson, N., & Johnson, J. (1979). Effect of the use of the vocational exploration group on career maturity. Journal of Counselling Psychology, 26, 238-270.

Zelkowitz, R.S. (1974). The construction and validation of a measure of vocational maturity for adult males. Doctoral dissertation, Teachers College, Columbia University, (University Microfilms No. 75-18, 456).

Development of an Air Force Training Decisions System

Sharon K. Garcia
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

I. Introduction

The management of technical training requires the coordination of efforts among numerous Air Force organizations. Significant roles in determining the nature of technical training are filled, for example, by the Air Force deputy chief of staff for manpower and personnel; by Air Force functional managers; and by agencies at the Air Force Manpower and Personnel Center in charge of job classification, assignments, and management of the on-the-job training programs. The Air Training Command at its headquarters, technical training centers, and other units plays a central role in policy as well as day-to-day training management. The headquarters of all major commands have a major voice in training decisions; and additional inputs are gathered from various ad hoc training study groups, AF laboratories, and research centers.

Because of the scope and complexity of the training and personnel systems, many decisions that impact on training are made, to some extent independently, by different management units responsible for different parts of the training and personnel systems. Conflicting goals are inevitable. As each unit attempts to optimize operations within its own area, the net result is competing objectives and total system suboptimization. Part of the problem is that relevant data for many training decisions are not available. For example, in practice, the amount of resident technical school training is largely determined by pre-defined budgets, and whatever content is not covered in the school is left to on-the-job training, with little data for assessing the impacts on OJT resources, costs, and capacities. In addition, alternative ways in which the overall training system might be restructured in concert with operational changes in personnel utilization are not considered. Hence, a more unified total systems approach to such problems with all relevant data considered was needed. In response to this need, Air Staff and the HQ Air Training Command requested the AF Human Resources Laboratory to conduct research in an attempt to refine current training decision procedures. The research, entitled the Training Decisions System (TDS), has as its objective the development of a computer-assisted decisions system to address the what, when, and where training decisions for any one specialty. What refers to what job tasks, skills, and knowledges should be emphasized in training. When refers to at what point in an airman's career training should be provided (e.g., upon entry, OJT, advanced training). Where, refers to the question of where training should be provided (e.g., resident school, field units, CDC, MAJCOM, other).

II. Development of the Training Decisions System (TDS)

The TDS will involve the development of four basic, user-friendly interactive subsystems. They are the Task Characteristics Subsystem (TCS), Field Utilization Subsystem (FUS), Resource/Cost Subsystem (RCS), and the Integration and Optimization Subsystem (IOS).

The Task Characteristics Subsystem (TCS) will be developed and used for identifying clusters of job tasks that can be appropriately trained as a unit, based on common skill and knowledge requirements and other shared characteristics (e.g., the probability of co-performance). These units will be known as Task Training Modules (TTMs) and will be used as the basic unit of analysis in the Training Decisions System. Once task training modules have been designed, a methodology will be developed for allocating these TTMs to the appropriate training settings or sites which may include initial skill resident training, on-the-job training, or correspondence courses.

The Field Utilization Subsystem (FUS) will be used to describe existing patterns of airman utilization in terms of jobs, training states, and major career paths. In addition, the FUS will attempt to define training/personnel assignment patterns that represent alternative approaches to training, assignment, and use of airmen in a particular specialty, based on management preferences. Both training content and job descriptions will be represented by collections of task training modules.

The Resource/Cost Subsystem (RCS) will be developed for estimating the costs and resources required for training different clusters of job tasks in alternative training settings; for estimating the training capacities of alternative training settings; and for developing summary estimates of costs, resources, and capacities needed for specified training alternatives.

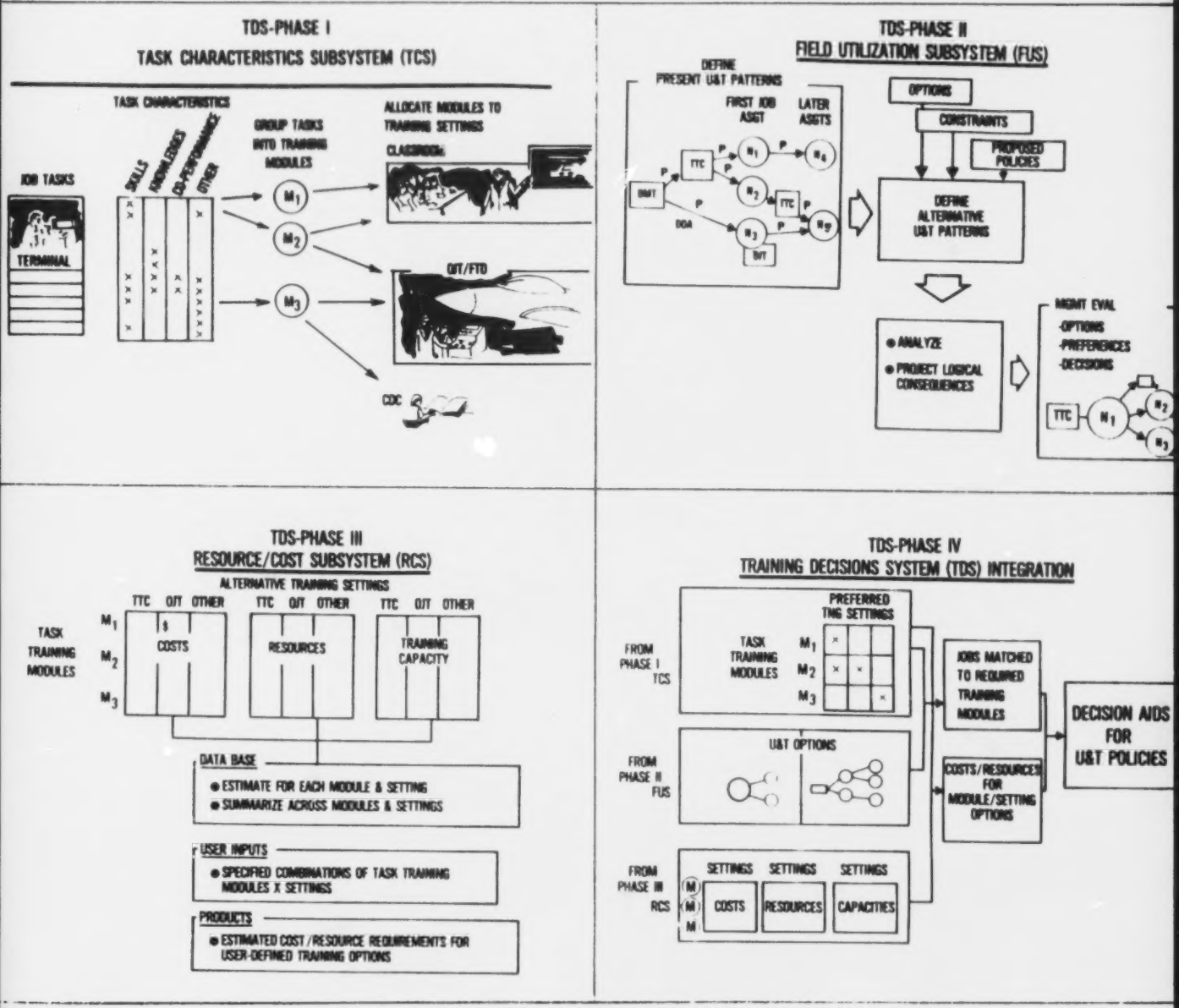
The Integration and Optimization Subsystem (IOS) will result in the integration of the three previously described subsystems and develop effective decision aids for AF technical training designers. The final product of the IOS will be the Training Decisions System; a user friendly software package, bringing together a complex of elements describing training, personnel, and cost factors, as well as management policy preferences. State-of-the-art decision model technologies will then be applied to these elements to answer "what if" questions for management, and to develop "optimal" training designs, for more cost-effective training decisions. A conceptual diagram of each of the TDS subsystems is provided in Figure 1.

Research and development of the TDS is a four-year contract effort. The prime contractor is McDonnell Douglas Astronautics. The contract began in Sept 83 and will be completed in Sept 87. Initial development of the TDS is being applied to four Air Force career ladders or occupations. They are Avionic Inertial and Radar Navigation Systems, Security/Law Enforcement, Aircraft Environmental Systems, and Electronic Computer and Switching Systems.

III. Conclusions

Once completed, the Training Decisions System will produce a training decisions system that will provide readily available, validated information to the Air Staff and user commands, especially Air Training Command, on costs and consequences of training decision alternatives under different constraints, costs, and personnel utilization patterns. The following benefits are anticipated from the implementation of such a system: (a) enhanced mission readiness through optimizing the match of technical training resources and overall operational demands, (b) increased training efficiency through optimizing the sequence and settings in which training occurs, (c) improved personnel utilization through development of methods for analyzing functional job patterns in relation to optimized training sequences, (d) increased cost

effectiveness of training through the formulation of training decisions based on explicit cost and resource consequences, and (e) reduction of excessive operational training commitments through more accurate estimation and analysis of unit capacity to train while meeting ongoing mission demands. In short, TDS will give managers the tool to plan for the best training for the dollar.



TDS CONCEPTUAL DIAGRAM

DEFINING TASK TRAINING MODULES: COPERFORMANCE CLUSTERING

Drs. B. M. Perrin, D. S. Vaughan, R. M. Yadrick,
& J. L. Mitchell
McDonnell Douglas Astronautics Company
Saint Louis, MO & San Antonio, TX

Training decision-making in the Air Force is a process of balancing, either explicitly or implicitly, a number of distinct and often conflicting considerations. Instructional effectiveness, manpower and personnel utilization plans, and financial factors must all be balanced in deciding who gets trained, when during their career, on what skills, and using which modes of instruction. Currently, Utilization and Training Workshops are used as a forum to weigh these considerations in determining training policy for an Air Force Specialty (AFS). Difficult decisions are made even more complex however, because the information available to these groups is often fragmentary, due to the number of distinct skills and knowledges in an AFS, the number of possible instructional modes, and the costs associated with training each skill using each mode.

The Training Decisions System (TDS) will be a means of bringing together information concerning each of these factors--instructional, personnel utilization, and financial--to aid Air Force managers in establishing training policy. The key elements of the TDS are the sets of skills and knowledges for which the relative instructional efficiencies of various training and utilization policies will be determined. These sets of skills and knowledges will be in the form of groups of Occupational Survey (OS) tasks, known as Task Training Modules (TTMs). Ideally, TTMs will be groups of OS tasks that are relatively homogeneous with respect to underlying skills and knowledges and that are relatively distinct from other groups of OS tasks (i.e., other TTMs); consequently, TTMS should capture efficiencies of training that might result from common training materials, content, equipment, and the like.

Additional training efficiencies accrue from training similar tasks together if these tasks are also coperformed (i.e., performed by the same personnel). Thus, TTMs should be composed of tasks which are similar and which are performed by the same personnel, if the information to be provided by the TDS is to be maximally useful. In a separate paper that follows in these proceedings (Yadrick, Vaughan, Perrin, and Mitchell, 1985), we have documented our method of obtaining expert judgments concerning task similarity. In this paper, we present procedures that may be used to hierarchically cluster tasks based on their reported coperformance in the OS. These statistical clusterings will then be compared to experts' groupings of the same tasks. Results are presented for two AFSs--328X4, Avionic Inertial and Radar Navigation Systems; and 811XX, Security, Law Enforcement, and Law Enforcement-Military Working Dog Qualified.

METHOD

Statistical clustering has been used for some time in military occupational analysis to aid job analysts in identifying job-types (i.e., groups of individuals performing similar tasks). The clustering technique that has been applied to this problem and is utilized in the Comprehensive Occupational Data Analysis Programs (CODAP) system is the average linkage clustering procedure (Ward, 1963). This procedure has performed well in empirical studies, as compared to other procedures reported in the statistical literature (Milligan, 1981; Mojena, 1977).

Figure 1 illustrates the relationship between case (or person) clustering, the normal application of CODAP to job-typing, and task clustering, the application to identifying tasks that are performed by the same personnel. (Note: Air Force job-typing normally uses a relative time spent measure for clustering cases rather than the performed/not performed dichotomy depicted in Figure 1; in this case, a number between 0 and 1 representing relative time spent on a task would replace the 1's in the figure.) While job-typing involves grouping persons who perform the same (or similar) sets of tasks, task clustering produces sets of tasks which are frequently coperformed.

Task clustering can be accomplished in CODAP by transposing the raw data file that is input to the system. That is, instead of task performance data for each person as represented on the left side of Figure 1, each record would consist of person performance data for each task (as represented on the right side of Figure 1). This transposed data would then be processed by CODAP, yielding a task coperformance cluster diagram.

Transposing the data input file has the effect of making the number of cases appear to the system as the number of tasks, and the number of tasks becomes the number of cases. This effect can have serious practical implications when the number of cases is large, as CODAP is limited in the number of "tasks" (transposed cases) it can process (the IBM version can handle up to 2000 tasks, and the UNIVAC version can process up to 1700, although a UNIVAC rewrite of CODAP will be able to process 3000 tasks).

FIGURE 1: A COMPARISON OF CASE CLUSTERING AND TASK CLUSTERING

JOB TYPING: CLUSTERING PERSONS WHO PERFORM THE SAME (OR SIMILAR) SETS OF TASKS

TTM CONSTRUCTION: CLUSTERING TASKS THAT ARE PERFORMED TOGETHER

9-3278

PERSONS		TASKS 1	2	3	4	5	6	7	8	9
JOB I	1	1	1	1						
	2	1	1	1						
JOB II	3	1	1	1	1	1	1			
	4	1	1	1	1	1	1			
JOB III	5				1	1	1			
	6				1	1	1			
JOB IV	7							1	1	1
	8							1	1	1

TASKS		PERSONS 1	2	3	4	5	6	7	8
TTM I	1	1	1	1	1				
	2	1	1	1	1				
	3	1	1	1	1				
TTM II	4			1	1	1	1		
	5			1	1	1	1		
	6			1	1	1	1		
TTM III	7							1	1
	8							1	1
	9							1	1

To avoid this limitation, one may compile a task similarity matrix external to CODAP, and then, use this matrix to cluster the task data. The performed/not performed similarity measure used in CODAP is as follows:

$$S_{ij} = ((N_{ij}/N_i) + (N_{ij}/N_j)) / 2$$

where N_{ij} is the number of persons performing both tasks i and j, N_i is the number of persons performing task i, and N_j is the number performing task j. In words, the similarity between pairs of tasks is the average of the two ratios of the number of persons performing both tasks divided by the number performing each task.

Both procedures--the transposition of the raw data and the computation of a similarity matrix--were used to analyze task coperformance in the 328X4 OS data to verify the methods. Because of the number of cases in the 811XX OS sample, in excess of 6000, a task similarity matrix computed outside of CODAP was used to cluster the data.

One measure of the homogeneity of the tasks taken to form a TTM is the between group similarity. Between group similarity is the average of the similarities of the tasks between the groups being merged to form a particular cluster. It is defined as follows:

$$BG_{ij} = \sum\sum S_{ij}/N_{ij}$$

where S_{ij} is the similarity between tasks i and j from the two groups to be merged and N_{ij} is the number of task comparisons between the groups.

Several statistics are available to compare clusterings or groupings of a single sample, based on a pairwise classification of cases for the two solutions. Table 1 illustrates the classification scheme, where each of the four cells specifies a type of agreement or disagreement between the solutions. For example, cell A indicates the number of pairs of cases grouped by both methods, while cell D is the number of pairs grouped separately by both solutions. Cell B and C indicate frequencies of disagreement in which pairs are grouped by one method, but not the other. Three comparison statistics can be defined in terms of these cell frequencies as follows:

Rand (1971): $(A + D) / (A + B + C + D)$

Jaccard (Downton & Brennan, 1980): $A / (A + B + C)$

Fowlkes & Mallows (1983): $A / \sqrt{(A + B)\ (A + C)}$

Table 1: A pairwise classification scheme used to compare two clustering solutions.

		Solution 1: pair in same cluster	Solution 1: pair not in same cluster
Solution 2	pair in same cluster	A	B
	pair not in same cluster	C	D

All three statistics yield a value of 1.00 when the two solutions agree perfectly, and all three have a lower bound of 0. The Rand statistic, despite being the most widely used, is often inflated by using pairs not classified together by either procedure (cell D) as reflective of solution consistency. Both the Jaccard and the Fowlkes & Mallows statistics were devised, in part, to overcome this problem. Sampling distributions are not available for these statistics; consequently, the numbers are not directly interpretable as indicating either agreement or disagreement between methods.

Two complications in comparing expert and statistical groupings of the tasks are worth noting at the outset. First, due to infrequent performance of some tasks within an AFS, the experts dropped some tasks from their groupings. While the number of deleted tasks was generally very small, one group of experts dropped 27 tasks from the 328X4 OS task list. When this complication occurred, these tasks were omitted from the analysis, and so, are not reflected in the comparison statistics.

The second complication resulted from experts placing the same task in two or more TTMs--an action they were directed to take if they believed it was necessary. Again, in absolute terms, this action was taken relatively infrequently, although one task statement was placed in five different TTMs by one group of experts. This complication was handled by counting each occurrence of the task individually; that is, one sorting by the experts may have agreed with the coperformance clustering (and be counted in cell A of Table 1), while a second sorting may have disagreed with the statistical clustering (and be counted in cell B or C).

RESULTS

Unlike case clustering to identify job-types which has more than 30 years of research and practice behind it, very little information exists to aid in identifying TTMs from a task coperformance cluster diagram. Rules-of-thumb which occupational analysts have adopted to narrow the search for important jobs may or may not be relevant to the search for TTMs. Particularly problemmatic is the determination of the number of TTMs. Since the clustering solution is hierarchical, TTMs of any degree of specificity can be identified.

An approach to this problem is to use the same type of heuristics a job analyst would use to interprete a case clustering. One of the authors of this paper, who is familiar with both occupational analysis and the 811XX career field, used this approach. He identified nineteen general task content areas (similar to the higher order job clusters in job analysis) and 67 TTMs within those general areas. The TTMs varied in size from 2 to 34 tasks and averaged just under 10 tasks per TTM. The between group similarities of the TTMs identified from the 811XX OS task coperformance clustering ranged from 26.75 to 91.20, and averaged 58.87.

While these TTMs, by definition, have the desirable characteristic of being composed of tasks which tend to be performed by the same personnel, the degree to which the tasks are similar in terms of skills and knowledges is not known. Some indication of the skill/knowledge homogeneity of the coperformance clusters can be obtained by comparing these results to those obtained from having experts group the tasks.

Two additional issues had to be addressed before the results of the expert and statistical clusterings could be directly compared. First, the number of statistical clusters had to be determined, since the comparison statistics are influenced by the specificity of the results. To promote comparability between the procedures, the number of statistical clusters was set equal to the number of expert task groupings for each comparison. The second issue dealt with how to select the statistical clusters; it was decided to use the task clusters which maximized between groups homogeneity. This purely statistical criterion was chosen because it permitted selection of different sets of TTMs with different degrees of specificity. This

approach also eliminated the variability that might have resulted from differences between analysts, had interpretations of the coperformance cluster diagram been used. Thus, the comparisons reflect the degree to which the statistical clusterings, without interpretation, capture expert task groupings.

Table 2 summarizes the results of the comparisons between the coperformance clustering produced by CODAP and two task groupings produced by separate groups of experts. Both the Jaccard and the Fowlkes & Mallows statistics are reported. The comparison between the two expert groupings in the 328X4 AFS produced the highest degree of convergence, substantially higher than that found between the 811XX expert groups. It should be noted, however, that the 328X4 groups consisted solely of technical trainers, while the 811XX groups contained both school and field personnel. The similarity of the backgrounds of the 328X4 experts may partially account for the greater convergence of their results.

The task coperformance clustering in the 811XX AFS matched the expert groupings as well as the experts' results matched each other. Additionally, the coperformance clustering agreed more closely with the experts' classifications in this specialty than in the 328X4 AFS. Again, this result may be due to the influence of the field personnel, whose perspective presumably is more attuned with performance factors.

CONCLUSIONS

The evidence concerning task coperformance clustering suggests that this procedure is workable and produces task groupings which are homogeneous with respect to content. Once computer software is developed, task coperformance clustering can be accomplished at relatively little additional expense, using existing OS data and CODAP clustering procedures. As this methodology develops, additional items may be added to the OS and computer routines derived to aid in the identification and interpretation of TTMS.

Table 2: Comparison of statistical and expert groupings of tasks in the 328X4 and the 811XX AFSs. The Jaccard and the Fowlkes & Mallows statistics are reported. Expert groupings I and II are results from independent efforts.

328X4 AFS	Expert Groupings - I	Expert Groupings - II	Coperformance Clusters
Expert Groupings - I	----	.271 .476	.087 .171
Expert Groupings - II		----	.121 .224
Coperformance Clusters			----

811XX AFS	Expert Groupings - I	Expert Groupings - II	Coperformance Clusters
Expert Groupings - I	----	.171 .293	.157 .314
Expert Groupings - II		----	.179 .350
Coperformance Clusters			----

The agreement between the statistical and expert groupings of the tasks, given that it was roughly equivalent to that between independent clusterings by experts, is encouraging. These results are even more promising considering that the task clusters were based solely on a statistical criterion, the between groups similarity. We would not recommend that this technique be directly implemented. Rather, an experienced analyst, working with one (or more) experts in the career field, should be used to identify and refine TTMs from a coperformance cluster diagram. This procedure would almost undoubtedly produce more homogeneous and useful TTMs.

One final word of caution is required. The skill/knowledge homogeneity of the task clusters produced by this methodology is still somewhat suspect, as the overall comparison statistics reported cannot specify the nature of the disagreement between results. Presumably, the experts' groupings might differ only according to subtle variations among tasks that are clustered by one group, but not the other. On the other hand, coperformance clusters, while exhibiting the same absolute number of differences, might include strikingly different types of tasks that are, nonetheless, coperformed. A more stringent validation effort is currently underway to assess this possibility.

REFERENCES

Downton, M., & Brennan, T. (1980, June). Comparison classifications: An evaluation of several coefficients of partition agreement. Paper presented at the meeting of the Classification Society, Boulder, Colorado.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings, (with comments and rejoinder). Journal of the American Statistical Association, 78, 553-584.

Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 16, 379-407.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. The Computer Journal, 20(4), 359-363.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.

Yadrick, R. M., Vaughan, D. S., Perrin, B. M., & Mitchell, J. L. (1985, October). Evaluating task training modules: SME clustering and comparisons. Paper prepared for the meeting of the Military Testing Association, San Diego, California.

EVALUATING TASK TRAINING MODULES: SME CLUSTERING AND COMPARISONS

DRS. R. M. YADRICK, D. S. VAUGHAN, B. M. PERRIN & J. L. MITCHELL

MCDONNELL DOUGLAS ASTRONAUTICS COMPANY
SAINT LOUIS, MO. & SAN ANTONIO, TX.

INTRODUCTION

One of the most straightforward ways to categorize information is simply to have people sort cards into piles. A single instance of the categories to be formed is printed on each card, and the piles that result constitute the categories. This procedure has a long history in psychology for research in cognitive modeling, and seemed suitable for use in forming Task Training Modules (TTMs) in support of the Training Decisions System (TDS) research (Garcia, 1985).

The approach was pilot tested at Scott AFB, with 811XX personnel (law enforcement) job incumbents serving as subjects, or Subject Matter Experts (SMEs). Cards were labeled with individual tasks from the most recent Occupational Survey (OS) on the 811XX AFS. Cards were initially grouped into "starter piles", in which all the tasks shared a common Specialty Training Standard (STS) paragraph reference. Starter piles provided SMEs with initial working units of manageable size, and also with a reasonable conceptual base (the STS paragraphs) from which to start.

SMEs were instructed to simply rearrange the starter piles to form groups of tasks that "should be trained together". Piles could contain many tasks, few tasks, or only a single task. Duplicate task cards could be placed in different piles. These rather flexible instructions and liberal sorting options were introduced in order to allow maximum expression of SME opinion and avoid forcing upon them any of our own notions about how TTMs should be formed.

There were two "passes" for each within-job sort. That is, SMEs initially rearranged the STS piles into their own piles. They then went through their own piles to make any needed changes. They worked at their own pace with no time constraints.

The results of this pilot test are reported elsewhere (Vaughan, Yadrick, Dunteman, and Clark, 1984) and need not be examined here in detail. Briefly, the resulting TTMs seemed like reasonable task groupings, and the overall card-sorting approach was feasible and deserved field testing. SMEs found it easy to work with the starter piles, performed conscientiously, and were satisfied with their own results.

In the present effort, the card-sorting method was field tested for 2 AFSs. The objectives of this project were to develop an operational system for TTM construction, and to get task clusters that could be compared to computer-generated clusters (Perrin, Vaughan, Yadrick, and Mitchell, 1985).

METHOD AND PROCEDURE

The card sorting process was modified and expanded for field testing purposes. Some of the modifications reflected policy decisions on our part (e.g., we decided not to have separate card sorts for each job in each specialty, but rather to have whole-AFS sorts), but most were minor changes designed to streamline the process.

The Security Police and Law Enforcement (811XX) card sorting was conducted at Lackland AFB, Texas, in May, 1985. The fourteen SMEs present represented a relatively even mix of the three 811XX shredouts, namely security police, law enforcement personnel, and military working dog (MWD) handlers. Technical trainers from ATC were represented in fairly equal proportion with operational personnel.

The SMEs were divided into four separate groups. These groups worked independently of the other groups, except during the final stage of the process. This stage will be described later. Again the groups were formed to provide the most even mix of SME backgrounds, shreds, etc., as possible. SMEs then received instructions and began sorting.

Two groups received starter piles in which all tasks shared a common STS reference, as in the pilot study. The other two groups received starter piles composed of tasks which had clustered together in the coperformance clustering process (Perrin, et al., 1985). The resulting TTMs could then be compared and evaluated. In addition, this would help mitigate any effect of starter pile content upon the final TTMs.

SMEs made three passes, all at their own pace. On the first pass, they oriented themselves to the exercise, examining piles for homogeneity of coperformance, skills and knowledges, combining and subdividing the piles as necessary. On the second pass, the newly created piles were recombined to refine the coperformance, knowledge and skill groupings. SMEs made further checks and refinements ("fine tuning") on the final pass. The results (TTMs) of each pass were recorded.

The four groups finished at very different times, as might have been expected. Two groups finished about halfway through the second day. Various groups dynamics were clearly observable, such as the domination of one group and of the sorting results by a single forceful member of the group.

Different groups worked as though they had quite different interpretations of the instructions. For example, one group arranged piles so as to train what they came to call a "super cop". They

carefully arranged sequences of TTMs so that, by receiving training on each full TTM in their prescribed sequence, the result would be an airman who could perform literally all the tasks and jobs in all the shreds of the specialty. Most groups, however, adopted very different and more realistic strategies.

The final phase of the process was conducted to reconcile the different sortings between groups. In each reconciliation, groups which had started with the same piles (e.g., the groups with STS starter piles) were teamed together.

Both reconciliation groups were each to provide a single set of "final" TTMs. Unfortunately, only one reconciliation group was actually able to finish. The other reconciliation group, made up of the two original groups which did not finish their own sortings in the first two days, were forced to rush through the reconciliation step.

The same essential process was carried out at Keesler AFB, with Avionic Inertial and Radar Navigation Systems (328X4) SMEs in August, 1985. There were, however, important differences. All of the SMEs available were instructors at the technical training school, and no operational people were present. Also, there were only enough SMEs to form two groups instead of four. As a result, the desired replication could not be done. Operational workers reviewed and refined the TTMs obtained at Keesler, although these refinements have not yet been examined in detail. The two 328X4 groups received essentially the same instructions a had the 811XX groups, although we stressed somewhat more the idea that TTMs should reflect common skills and knowledges required to do particular jobs. This change was made mainly to avoid any misinterpretation of instructions.

RESULTS

In the 811XX AFS, the reconciliation effort resulted in 69 TTMs from the groups that had coperformance cluster starter piles. Of these, only four TTMs contained a single task, and 52 contained five or more tasks. The largest contained 42 tasks. These were sorted from 19 starter piles containing a total of 666 tasks.

The other reconciliation effort (groups with the STS starter piles) resulted in 65 TTMs formed from 39 started piles. Only one contained a single task. The largest contained 31 tasks.

Although these results appear to be in reasonably good agreement, the similarities are only superficial. As reported in another paper in this session (Perrin, et al. 1985), both the Jaccard index (Downton & Brennan, 1980) and that from Fowlkes and Mallows (1983) were relatively low (0.171 and 0.293, respectively) showing relatively poor agreement between the two groups in assigning tasks to TTMs. Indeed, both indices were higher (0.179 and 0.350, respectively) and showed greater agreement between the STS starter pile reconciliation and the pure coperformance clustering than between the two card sorting reconciliation groups.

In the 328X4 AFS, the group which started with STS piles formed 140 TTMs from 31 original piles containing 778 tasks. For the most part, these were small. Eighty-six contained fewer than 5 tasks, and many of these contained only one task. However, two were especially large, containing 78 and 48 tasks, respectively.

The group that started with coperformance piles formed 33 TTMs from 58 original piles. Eleven of these contained over thirty tasks, and two contained over eighty tasks.

Although there appears to be considerable disagreement between the groups, it is apparently superficial. Table 1 reports Jaccard and Fowlkes-Mallows indices for several comparisons of these card sorts

---

Table 1: Comparison of SME card sorts and task coperformance statistical clustering in the 328X4 AFS*

	Reconciliation Card-Sort	Card-Sort From Coperformance Starter Piles	Card-Sort From STS Starter Piles	Task Coperformance Clusters
Reconciliation Card-Sort	----	.326 .520	.637 .787	.127 .244
Card-Sort From Coperformance Starter Piles		----	.271 .476	.087 .171
Card-Sort From STS Starter Piles			----	.121 .224
Task Coperformance Clusters				----

*For each table entry, the Jaccard statistic is on top, and the Fowlkes-Mallows statistic is on the bottom.

---

with each other, with the later reconciliation sort done between the two groups, and with coperformance clustering. In general, the two groups actually agreed relatively well. The obvious conclusion is that the 140 TTMs formed by the one group were generally subgroupings of the 33 TTMs formed by the other. Indeed, the reconciliation data indicate fairly good agreement with both groups (see Table 1). Also, reconciliation

resulted in 75 TTMs, roughly halfway between the numbers of TTMs produced by the groups separately. It would seem that they found compromise easy to make. Their TTMs, however, do not agree very well with the computer-generated coperformance task clusters. As Table 1 shows, both the Jaccard and Fowlkes-Mallows statistics were less than .25 for all such comparisons.

DISCUSSION

The results reported here are tentative. Final TTMs have not yet been produced for either AFS studied. The 328X4 TTMs must be replicated with additional field data. Nonetheless, even at this point it appears that neither the clusters derived from task coperformance clustering nor those constructed by a single group of SMEs would yield stable TTMs. Instead, some combination of methods will be required that allows tentative TTMs to be crossvalidated and refined by representative groups of field experts. Once a method is developed that produces final, stable TTMs for each specialty, these TTMs can be used as criteria against which more economical clustering methods can be validated.

REFERENCES

Downton, M., & Brennan, T. (1980, June). Comparing classifications: An evaluation of several coefficients of partition agreement. Paper presented at the meeting of the Classification Society, Boulder, Colorado.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings, (with comments and rejoinder). Journal of the American Statistical Association, 78, 553-584.

Garcia, S. (1985, October). Overview of the Training Decisions System. Paper prepared for the meeting of the Military Testing Association, San Diego, California.

Perrin, B. M., Vaughan, D. S., Yadrick, R. M., & Mitchell, J. L. (1985, October). Defining task training modules: Coperformance Clustering. Paper prepared for the meeting of the Military Testing Association, San Diego, California.

Vaughan, D. S., Yadrick, R. M., Dunteman, G. H., & Clark, B. L. (1984). Feasibility of task training module construction methods and preliminary task characteristics subsystem project design. Saint Louis, MO: McDonnell Douglas Corporation.

NONHIERARCHICAL CLUSTERING OF AIR FORCE JOBS AND TASKS

Jimmy L. Mitchell, PhD
McDonnell Douglas Astronautics Co.
San Antonio, Texas

William J. Phalen
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

INTRODUCTION

One of the problems with the typical occupational analysis project is the limitation of coverage of job groups in the normal Comprehensive Occupational Data Analysis Programs (CODAP) hierarchical clustering solution. In an operational study, the groups of minimally acceptable size existing at any stage of the clustering, as reflected on the diagram of the occupation, will not include all the cases, except at a very low stage, where between-group overlap values are also very low (see Figure 1). This is a function of both the degree of homogeneity of jobs within the specialty and the order in which the cases enter the initial groups. It is not unusual to have 5 to 20 percent of the cases in a study excluded from the final job groups identified in an occupational survey report (OSR).

While the identification of the job groups in an occupation or Air Force Specialty (AFS) has considerable utility in the personnel classification and training system, the final job descriptions for such groups have some inherent limitations (Carpenter, 1974; Pass & Robertson, 1978). The hierarchical clustering procedure normally used is an iterative process whereby, once a case is clustered in a group, it is no longer considered in terms of its similarity of tasks performed or relative time spent performing against later-formed groups. Some researchers have proposed that other clustering methodologies or other similarity measures be used (Pass, 1980). No one approach, however, has emerged as an optimum clustering alternative.

In the present study, a nonhierarchical clustering methodology was used to refine the job types identified in the normal CODAP process. The usual hierarchical clustering procedure was used to identify the "seed" groups and an iterative, nonhierarchical clustering method was added to refine these groups. The expectation was that a combination of the two methods would provide superior results in terms of group membership, percent of cases accounted for, and realism of the resulting job descriptions.

METHODOLOGY

A runstream of existing CODAP programs has been put together to give the CODAP system a nonhierarchical capability (Datko, 1985; Phalen, Weissmuller, & Staley, 1985). This new procedure was developed to facilitate analysis of various types of officer job rating scales (relative time spent versus part of the job, complexity, etc.), and has already proven very useful in that context.

If a set of CODAP-generated job descriptions is input to this nonhierarchical clustering program as "seed" profiles, an unlimited number of cases, as represented by their individual job descriptions, can be classified into homogeneous groups, according to which one of the "seed" profiles each case resembles most closely. Multiple iterations can be run to ensure that each case has the opportunity to switch to a more compatible group, or to switch to a group whose homogeneity is increased the most by classifying the case in it, until the number of reclassifications has reached a minimum. The reclassification of cases, previously grouped by a hierarchical clustering, is a way to improve within-group homogeneity and between-group differences (Phalen et al., 1985). Also possible in the system is multiple group membership by a case or group of cases.

While the membership of the nonhierarchically clustered job types will overlap to a very substantial degree with those produced by the normal CODAP clustering, this new procedure permits the "isolates" or unclustered cases to be included in the covered sample. In addition, the resulting job descriptions should be considerably more homogeneous in terms of both the "core" tasks of the group and the relative time spent on those tasks. Thus, the relative worth of the new procedure could be assessed by examining the increase in the homogeneity of the groups. Such a procedure, then, should result in a clearer definition of the variety of jobs within a specialty, and a more realistic determination of the training requirements for each such job.

SAMPLE

CODAP case data were readily available for a sample of first-enlistment Security Policemen (AFSC 811X0) who were being studied in another line of research (Perrin, Vaughan, Yadrick, & Mitchell, 1985). The sample included 3,302 first-enlistment individuals from the most recent Security Police occupational analysis study (Alton, 1984). A separate first-enlistment diagram was analyzed and 82 811XX first-term seed job groups were identified which had reasonable internal homogeneity (overlap between combining groups ≈ 35 and overlap within combined groups ≈ 50). Group size ranged from 3 to 210; the average group size was 27. Some very small groups were included which were less than the starter group size of 10; these smaller groups were identified by closely examining several small, undefined heterogeneous groups with mixed job titles and low within-group overlap values (such as Customs or Kennel Support). The reason for including these very small groups was to see if they would gain sufficient additional members in the nonhierarchical process to become legitimate job groups.

The 82 groups accounted for only 67 percent of the cases in the sample, which illustrates the coverage problem discussed earlier. Each group was named and related groups were given overlapping component names to facilitate analysis of their relationships.

RESULTS

The initial nonhierarchical grouping yielded a marked reduction in the number of cases not classified into the 82 seed job types. Only five cases remained unclassified for a 99.8 percent coverage (see Figure 2). In succeeding iterations, the number of unclassified cases increased slightly, but even at the sixth iteration, the proportion of cases covered was greater than 99 percent. Thus, it would appear that almost all of the unaccounted-for cases in the original clustering were reasonably similar to the major job types identified in the initial analysis. The question now became one of assessing the impact of adding these previously unaccounted-for cases to the groups.

One way to address this issue was to examine the within-group overlap values, averaged across all 82 groups. For the initial iteration of the nonhierarchical clustering process, the average within-group overlap was 48.23. The within-group standard deviation averaged across all groups was 10.68. For the second run, the mean within-group overlap value was 48.65 (vice 48.23) with an average S.D. of 8.81 (vice 10.68). Thus, this second iteration had little if any impact on the average within-group overlap but resulted in a substantial decrease of the within-group variance. This change in values indicated that the new job group descriptions were considerably more homogeneous. This is, however, a summary statistic and, in order to demonstrate a meaningful impact on the job-typing process, we must examine how the addition of the unclassified cases changed individual job groups. Some representative results were selected to illustrate several observed trends.

Three related Law Enforcement (LE) jobs identified in the original hierarchical clustering study included: LE Desk Sergeants (Grp 594), LE Patrolmen (Grp 785), and a group which performs both as Desk Sergeants and Patrolmen (Grp 921). In the original study, group size for these groups was 15, 12, and 90, respectively (see Figure 3). In the initial iteration of the nonhierarchical clustering process, both the Desk Sergeant and Patrolmen groups more than doubled in size, whereas the mixed group dropped in membership. It is reasonable to assume that the 22 members lost from the combined LE Desk Sgt/Patrol group were those performing more Desk Sergeant functions, since 22 new members appeared in the Desk Sergeant group (Grp 594). Note that once these 22 new cases were added, the Desk Sergeant group membership essentially stabilized--the number of cases, the mean within-group overlap, and the standard deviation remained about the same for iterations 1 through 6. Note also that the standard deviation for this larger group (N = 37) dropped considerably (from 9.2 to 7.3), which demonstrates the development of a more homogeneous group.

The composite LE Desk Sgt/Patrolmen group (Grp 921) attritted more of its membership with each iteration of the process. Its mean within-group overlap dropped slightly at each stage and its SD increased, indicating that the remaining cases were more heterogeneous than in the original seed group. Presumably, if additional iterations had been run, this group would essentially disappear.

The LE Patrolmen (Grp 785), on the other hand, first grew in membership and then shrank, while its mean within-group overlap steadily increased and its SD dropped. The group gained in membership from other groups not included in this comparison (there is a composite LE Patrolmen/Entry Controller group, and other possible contributors). The point is that even at the sixth iteration, this group had not completely stabilized, and additional runs might be needed to maximize the within- group overlap and minimize the within-group variance.

A set of three related Alert Area Security groups illustrates some additional trends (see Figure 4). All three groups involve tasks being performed by Security Policemen in controlling access to and guarding alert aircraft. Grp 283 appears to be the most meaningful of the three groups, in that its membership continued to expand through all six iterations, whereas the other two groups fluctuated. We might need to run additional iterations to see how large this group would become before the mean overlap is maximized and the S.D. is minimized. For the other two groups, we need to examine their job descriptions to determine how they differ from Grp 283 (that is, what makes them distinct groups). Only then can a judgment be made whether these two small, low-overlap groups should be retained.

DISCUSSION

Preliminary results strongly suggest that the nonhierarchical clustering process has considerable potential in refining the existing occupational analysis process. The drastic drop in the number of unclustered or unaccounted-for cases represents the greatest benefit, since the group job descriptions resulting from increased sample size should be much more stable. We need, however, to extend our analysis to include study of such variables as the core tasks for each group and the amount of core job time accounted for by such core tasks. With lower within-group variance and somewhat increased within-group overlap, the set of core tasks should account for a significantly greater proportion of total work time. Such a finding would have important implications for the identification of training requirements within a specialty, both for resident programs and on-the-job training programs. In addition, the more discretely defined job groups resulting from this type of analysis should have considerably more construct and content validity in the eyes of other occupational analysis users, such as functional managers in the manpower and personnel communities.

Clearly, further work needs to be done to define the proper techniques for using this type of system. Data displays should be developed which will permit an analyst to track what is happening to each group across iterations. Also needed is a method of setting aside stabile groups at the end of each iteration and then continuing to run the program on the remaining groups. The analysis process will have to be sufficiently flexible to permit multiple runs of the later iterations in order to optimize both the mean group overlap and standard deviation values. Such developments are planned over the next few months.

The system could also be used to cluster a transposed task-by-person matrix (Perrin et al., op cit) to refine Task Training Modules (TTMs) from coperformance data. Such an application would have considerable advantages in terms of reduced cost and multiple TTM membership for technical tasks that apply to more than one module.

ACKNOWLEDGEMENTS

Mr William J. Phalen, AFHRL/MOMM, developed the concept and wrote the specifications for the nonhierarchical clustering algorithm used in this research. Sgt Kevin Kumba, AFHRL Technical Services Division, wrote the computer programs involved, and Mr Lou Datko, AFHRL/MODJ (now of the USAF Manpower and Personnel Center, Randolph AFB, TX), contributed substantially to their refinement during initial testing.

REFERENCES

Alton, R.L. (1984, November). Occupational survey report, security police career ladders (AFSCs 811X0, 811X2, and 811X2A). Randolph Air Force Base, TX: USAF Occuational Measurement Center, Occupational Analysis Program.

Carpenter, J.B. (1974, March). Sensitivity of group job descriptions to possible inaccuracies in individual job descriptions (AFHRL-TR-74-6, AD-778 839). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.

Datko, L.M. (in preparation). Rating scales for officer occupational analysis. Brooks AFB, TX: Air Force Human Resources Laboratory. Draft Technical Report.

Pass, J. (1980, May). An empirical comparison of proximity measures for CODAP cluster analysis. Proceedings of the Third International Occupational Analysts Workshop. Randolph AFB, TX: USAF Occupational Measurement Center.

Pass, J.J., & Robertson, D.W. (1978). Sample size and stability of task analysis inventory response scales. Proceedings of the 20th Annual Conference of the Military Testing Association. Oklahoma City: U.S. Coast Guard Institute.

Perrin, B.M., Vaughan, D.S., Yadrick, R.M., & Mitchell, J.L. (1985, October). Defining task training modules: Coperformance clustering. Proceedings of the 27th Annual Conference of the Military Testing Association. San Diego, CA: Naval Personnel Research & Development Center.

Phalen, W.J., Weissmuller, J.J., & Staley, M.R. (1985, May). Advanced CODAP: New analysis capabilities. Proceedings of the Fifth International Occupational Analysts Workshop. Randolph AFB, TX: USAF Occupational Measurement Center.

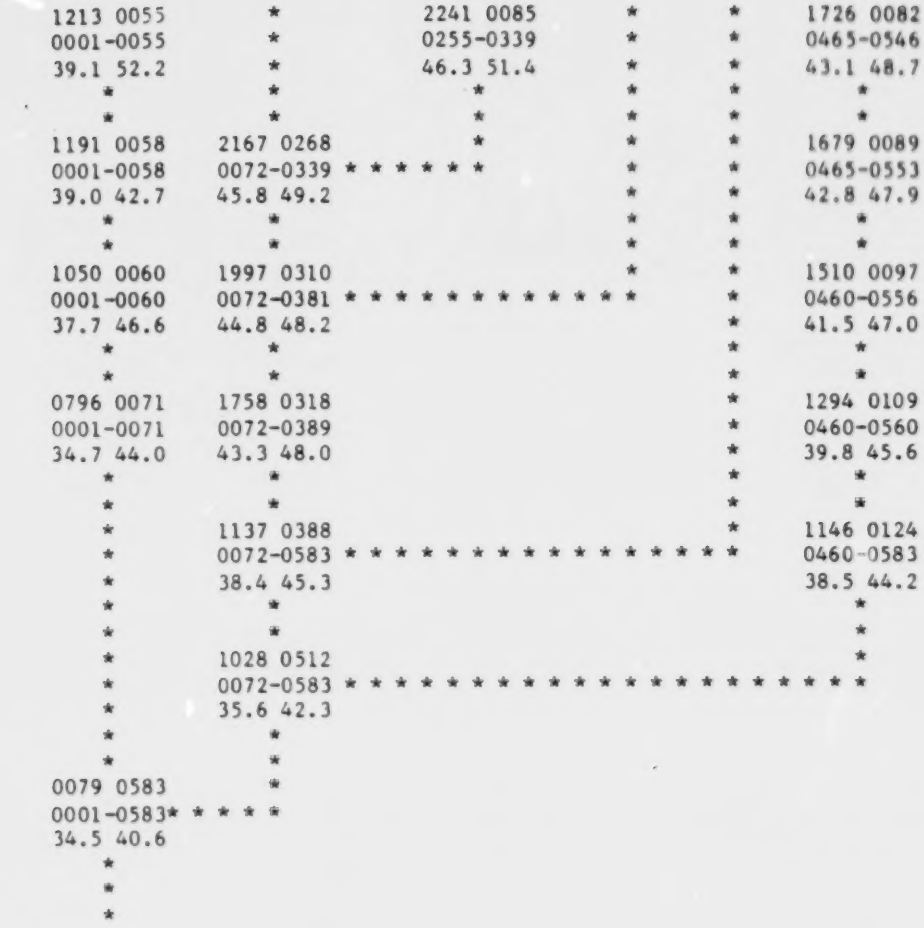


Figure 1. Clustering of Security Police Time-Spent Ratings - 811XX

Iteration	No. of Cases Classified	Number Unclassified	Percent of Coverage
0 (input)	2238	1065	67.7%
1	3298	5	99.8%
2	3292	11	99.6%
3	3285	18	99.4%
4	3281	22	99.3%
5	3279	24	99.3%
6	3278	25	99.2%

Figure 2. Changes in the Number of Cases Classified

Group	Variable	Seed	Iteration 1	2	3	4	5	6
Grp 594 LE Desk Sergeant								
	Group Size	15	37	34	33	32	32	33
	No. Lost	-	-	3	2	1	0	0
	No. Gained	-	22	0	1	0	0	1
	Mean	-	44.6	42.5	42.7	42.4	42.5	42.6
	S.D.	-	9.2	7.4	7.6	7.4	7.3	7.3
Grp 921 LE Dsk Sgt/Patrol								
	Group Size	90	68	38	20	14	10	5
	No. Lost	-	22	30	18	7	4	5
	No. Gained	-	0	0	0	1	0	0
	Mean	-	57.4	56.2	56.9	55.3	54.4	49.2
	S.D.	-	7.2	7.4	8.2	9.9	10.5	10.9
Grp 785 LE Patrolmen								
	Group Size	12	38	55	90	84	67	51
	No. Lost	-	-	6	12	29	25	20
	No. Gained	-	26	23	47	23	8	4
	Mean	-	51.0	52.1	55.5	57.2	57.6	57.9
	S.D.	-	10.9	7.8	7.4	7.2	6.9	5.9

Figure 3. Changes in Several Law Enforcement Groups

Group	Variable	Seed	Iteration 1	2	3	4	5	6
Grp 283 Alert Area Sentry/Entry Controller								
	Group Size	54	73	113	170	231	277	279
	No. Lost	-	-	4	5	10	17	34
	No. Gained	-	19	44	62	71	63	36
	Mean	-	44.2	46.2	47.9	48.3	49.1	49.8
	S.D.	-	10.7	9.9	9.7	9.5	9.6	9.4
Grp 361 (Same Title)								
	Group Size	14	45	46	34	33	31	30
	No. Lost	-	-	9	14	4	3	1
	No. Gained	-	31	10	2	3	1	0
	Mean	-	36.4	33.7	31.0	32.4	32.1	32.1
	S.D.	-	14.9	11.6	10.4	8.9	8.4	8.1
Grp 470 (Same Title)								
	Group Size	11	18	14	14	14	13	11
	No. Lost	-	-	4	0	0	1	2
	No. Gained	-	7	0	0	0	0	0
	Mean	-	41.4	39.3	40.7	40.7	40.1	38.6
	S.D.	-	17.7	11.7	10.3	10.3	10.5	10.3

Figure 4. Changes in Related Alert Area Security Jobs

On the Study of Differential Item Performance without IRT

Paul W. Holland
Educational Testing Service
Princeton, New Jersey 08541

1. INTRODUCTION

The problem of identifying items for which the performance of certain subpopulations -- often women and minorities -- is unusual and out of line with their performance on other items or test results has a substantial history. The book by Berk (1982) summarizes the state of the art as of 1980 and the work of Lord (1980), Scheuneman (1979), Shepard, et.al. (1981), among others are relevant. From a statistical point of view, modern methods for the analysis of multi-way contingency tables seem particularly appropriate to this problem and some suggestions for their use have been made, (Marascuillo and Slaughter, 1981). In this spirit, the present paper proposes the well-known method of Mantel and Haenszel (1959) for the analysis of 2×2×K contingency tables as an easily implemented, powerful technique for the measurement of the degree to which two subpopulations of examinees perform differently on a given test item. Modern references to the Mantel-Haenszel procedure include Breslow (1981), Hauck (1979), and Breslow and Liang (1982). The basis for the use of the Mantel-Haenszel (herein MH) procedure in the study of differential item performance is the fundamental notion of the need to compare comparable people when examining the relative performance of two groups of examinees on an item. This is the problem of matching and is discussed in section 2. Section 3 gives the relevant facts about the MH procedure while section 4 discusses various aspects of the MH procedure and related methods in the context of measuring differential item performance.

2. MATCHING VERSUS CONTROLLING FOR ABILITY

The need to "control for ability" is well established in the differential item performance literature. It is the fundamental basis for the proposed use of item response theory methods to study "item bias." Other methods, such as those of Scheuneman (1979), use test performance as a proxy for ability. The "delta-plot method," Angoff (1982), controls for ability indirectly by concentrating attention on the covariance between the item difficulty indices for the two groups rather than on their respective mean values.

In my opinion, the "need to control for ability" is an inadequate way to express a more fundamental idea. When we compare two subpopulations on any criterion, it is always important to be sure that only comparable members of the two groups are being compared. What constitutes comparability will depend on the problem at hand. In the study of differential item performance we are interested in learning something about a test item and how members of one subgroup (the "focal group") might react differently to it than do the members of another subgroup (the "reference group"). If our criterion is performance (i.e., right or wrong on the test item) then it is improper to compare the performance of reference and focal group members who differ in significant and

measurable ways that are related to their performance on the item. Differential item performance means differences in performance on an item between focal and reference group members that is attributable to characteristics of the item and not to differences in characteristics of the groups of examinees.

When we confound both examinee characteristics and item characteristics and simply look at differences in the performance on an item of reference and focal group members we are measuring what is called <u>impact</u> rather than differential item performance. For example, comparing the proportion of reference and focal group members who give correct answers to a given item is a measure of the item's impact on the focal group relative to the reference group. In measuring differential item performance members of the reference and focal groups are first divided into sets of examinees who are <u>matched</u> on relevant criteria before their performance on the item is compared. Examples of relevant matching criteria are: scores on related tests, schooling measures, and other group membership. In many practical settings, matching will be done on related test scores since these are both available and accurately measured.

The 2×2×K Table: For a given item, say item j, the data from the i^{th} matched group of reference and focal group members can be arranged as a 2×2 table:

	Right on item j	Wrong on item j	
Reference	A_i	B_i	n_{Ri}
Focal	C_i	D_i	n_{Fi}
Total	R_i	W_i	n_{+i}

(1)

For i=1,..., K = number of matched groups. In (1) A_i denotes the number of reference group members in the i^{th} matched group who answered item j correctly. B_i, C_i, and D_i have corresponding interpretations. n_{Ri} and n_{Fi} denote the number of reference and focal group members, respectively, in the i^{th} matched group, while n_{+i} denotes the total number in the i^{th} matched group of examinees. R_i and W_i denote the number in the ith matched group who get the item right and wrong, respectively. Considered together these K 2×2 tables form one big 2×2×K table. There is one such 2x2xK table for each item being considered. It is worth emphasizing that once the criteria for matching have been selected, the 2×2×K table of data can be formed from samples of data from the reference and focal group members. It should also be emphasized that the choice of matching variables is important and will depend on the availablity, amount, and accuracy of data as well as on its relevance to item performance.

3. THE MANTEL-HAENSZEL PROCEDURE

In the i^{th} matched group, the odds that a reference group member gets item j correct is A_i/B_i, while the corresponding odds for a focal group member is C_i/D_i. The MH procedure measures the advantage (or disadvantage) on item j that reference group members have relative to their matched focal group colleagues by the ratio of these two odds. This gives us the odds-ratio estimate

$$\hat{\alpha}_i = \frac{A_i}{B_i} \Big/ \frac{C_i}{D_i} = \frac{A_i\ D_i}{B_i\ C_i} \ . \qquad (2)$$

The $\hat{\alpha}_i$ estimate a population cross-product-(or odds-) ratio, α_i, for the i^{th} matched group.

The Mantel-Haenszel common-odds-ratio estimate is a weighted average of the $\hat{\alpha}_i$ that uses the following weighted formula:

$$\hat{\alpha}_{MH} = \frac{\Sigma\ \omega_i\ \hat{\alpha}_i}{\Sigma\ \omega_i} \ , \qquad (3)$$

where

$$\omega_i = \frac{B_i\ C_i}{n_{+i}} \ . \qquad (4)$$

Substituting (4) into (3) yields the usual formula for $\hat{\alpha}_{MH}$:

$$\hat{\alpha}_{MH} = \frac{\Sigma\ A_i\ D_i/n_{+i}}{\Sigma\ B_i\ C_i/n_{+i}} \ . \qquad (5)$$

The Mantel-Haenszel estimate, $\hat{\alpha}_{MH}$, is the average factor by which the likelihood that a reference group member gets item j correct exceeds the corresponding likelihood for comparable focal group members. (Likelihood is measured by the odds of getting item j correct). For example, if $\hat{\alpha}_{MH} = 1$ then reference and focal group members are, averaging across all the matched groups, equally likely to be correct on the item. When $\hat{\alpha}_{MH} > 1$ then the reference group has the advantage whereas when $\hat{\alpha}_{MH} < 1$ the focal group has the advantage.

Associated with the estimate $\hat{\alpha}_{MH}$ is a one-degree-of-freedom chi-square test of the hypothesis that all of the population cross-product ratios in all of the 2×2 layers of the 2×2×K table are unity (i.e., $\alpha_i = 1$ all i). This test is given by the formula:

$$\chi^2_{MH} = \frac{(|\Sigma_i\ A_i - \Sigma_i\ \mu_i| - \frac{1}{2})^2}{\Sigma_i\ \sigma_i^2} \qquad (6)$$

where

$$\mu_i = E(A_i|\alpha_i=1) = \frac{n_{Ri}\ R_i}{n_{+i}} \qquad (7)$$

and

$$\sigma_i^2 = Var(A_i|\alpha_i=1) = \frac{n_{Ri}\ n_{Fi}\ R_i\ W_i}{(n_{+i})^2(n_{+i}-1)}\ . \qquad (8)$$

The χ^2_{MH} from (6) will be large if $\hat{\alpha}_{MH}$ differs from 1.0 significantly in either direction. Thus, this test will detect differential item performance that favors either the reference or the focal group.

4. DISCUSSION

The MH procedure is closely related to log-linear model procedures for estimating a constant two-way interaction across a series of 2×2 tables (see Bishop, Fienberg, and Holland, 1975). In practical terms, $\hat{\alpha}_{MH}$ is usually nearly identical to estimates of the common-odds-ratio that involve complicated iterative procedures. While the formula for $\hat{\alpha}_{MH}$ is a simple weighted average of the sample odds-ratio $\hat{\alpha}_i$, it has been shown (Breslow, 1981) that, over the range of values relevant to this application of the MH procedure, $\hat{\alpha}_{MH}$ is nearly optimal as an estimator. In other words, no other estimate of the common-odds-ratio can have a substantially smaller variance. The chi-square test based on χ^2_{MH} is of high power because it is concentrated into a single degree of freedom rather than dissipated across several degrees of freedom.

If there is more than one pair of groups that could serve as the reference and focal group in an analysis then values for $\hat{\alpha}_{MH}$ and χ^2_{MH} can be computed for all such pairings.

The parameter $\Delta = -2.35\ \ln(\alpha)$ is (approximately) in the scale of differences in delta-units of difficulty where delta-units are those used by ETS in their normal item analysis procedures. This transformation can be used to put $\hat{\alpha}_{MH}$ values into units which are familiar to those who use the delta-scale in test construction and analysis:

$$\hat{\Delta}_{MH} = -2.35\ \ln(\hat{\alpha}_{MH}). \qquad (9)$$

Thus, $\hat{\Delta}_{MH} = -1.0$ means that the focal group found the item one delta-unit harder

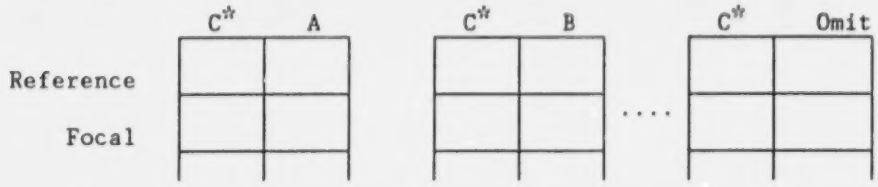
than did comparable members of the reference group. The parameter $\hat{\Delta}_{MH}$ is similar to an average shift to the right of $-\Delta_{MH}$ in the ICC of the focal group relative to the ICC of the reference group in an IRT model (as estimated by LOGIST).

The MH procedure can easily be expanded to include an analysis of distractor choice for multiple choice tests. For five-choice responses the 2x2 table in (1) is replaced by the followin 2x6 table

Response on item j

	A	B	C*	D	E	Omit	Total
Reference							n_{Ri}
Focal							n_{Fi}
Total							n_{+i}

C^* is correct answer, for example. Then the MH procedure is applied to the five 2×2 tables formed by juxtaposing the column for the correct answer with a column for one of the five ways of producing incorrect answers. E.g.,



This yields five MH cross-product estimates and five chi-square tests for each item. In some cases these may be used to see if a significant value of $\hat{\alpha}_{MH}$ is due to a single type of incorrect answer.

There are a number of important research issues that need to be addressed in the use of the MH procedure in the study of differential item performance. What aspects of the criteria for matching examinees seriously affects $\hat{\alpha}_{MH}$ in practical settings -- the reliability of the criteria, the fineness of the matching, the use of other examinee attributes, etc.? How stable are the values of $\hat{\alpha}_{MH}$ across different examinee populations? What are the relationships between the values of $\hat{\alpha}_{MH}$ and other statistical indices used to construct tests -- i.e., difficulty and discrimination? How should values of $\hat{\alpha}_{MH}$ for several pairs of reference and focal groups be combined for the same test item?

The MH procedure promised to be a relatively inexpensive and yet statistically powerful technique for identifying test questions that are potentially "biased" or unfair in some way to identified groups of examinees. ETS is currently embarked on a variety of research projects to see how to best use this tool for such purposes.

References

Angoff, W. (1982) "Use of difficulty and discrimination indices for detecting item bias" in Berk, R. (ed.) Handbook of Methods for Detecting Test Bias. Baltimore and London: Johns Hopkins University Press.

Berk, R. (Ed.) (1982) Handbook of Methods for Detecting Test Bias. Baltimore and London: Johns Hopkins University Press.

Breslow, N. (1981) Odds ratio estimates when the data are sparse. Biometrika, 68, 73-84.

Breslow, N. and Liang, K. (1982) The variance of the Mantel-Haenszel estimator. Biometrics, 38, 943-952.

Hauck, W. (1979) The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. Biometrics, 35, 817-819.

Lord, F. (1980) Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Marascuillo, L. and Slaughter, R. (1981) Statistical procedures for identifying possible sources of item bias based on chi-square statistics. Journal of Educational Measurement, 18, 229-248.

Scheuneman, J. (1979) A new method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L., Camilli, G., and Averill, M. (1981) Comparisons of six procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

BEST COPY AVAILABLE

REVIEWING AN ITEM POOL FOR ITS SENSITIVITY TO THE CONCERNS OF WOMEN AND MINORITIES: PROCESS AND OUTCOMES

Susan Wilson Kershaw
Howard Wainer
Educational Testing Service
Princeton, New Jersey

Why a Sensitivity Review

An item pool for a computerized adaptive version of the ASVAB has been developed and must be subjected to numerous editing and review processes before being made operational. Among these processes is a sensitivity review of the entire item pool for material that may be potentially offensive to minority groups and/or women. During a sensitivity review, all test items are screened by trained sensitivity reviewers to ensure that the test is as free as possible from perceived bias and offensiveness. In order for each examinee to perform at his or her optimum level of ability, it is necessary to eliminate any material that may convey negative or distracting messages to a particular subgroup of the test-taking population. This type of review is particularly important in the development of the CAT-ASVAB since the test-taking population will be drawn broadly from many cultural and socio-economic backgrounds and because it is difficult to forsee what particular combination of items will be presented to a given examinee.

The Sensitivity Review Process

ETS Objectives

The overall objective of the sensitivity review process at ETS is to eliminate any material from tests that may be potentially offensive or inappropriate for identifiable subgroups of the test-taking population, or that reinforces negative attitudes toward these subgroups. The ETS test evaluation criteria consist of a general set of criteria that can be applied to all people, and specific criteria that are particularly relevant for five subgroups--Asian Americans, Black Americans, Hispanic Americans, Native Americans, and women.

Test Evaluation Critieria

The following test evaluation criteria have been excerpted from ETS Test Sensitivity Review Process (Hunter & Slaughter, 1980). For a more detailed description of the sensitivity review process, refer to their document.

A. Definitions
Group reference items reflect the multicultural nature of our society. There are two basic classes of subgroup reference items: representational items and substantive items.

1. Representational items are those in which references are made to minorities and women, but where the subject matter content of the test is intended to measure factors unrelated to such groups. Reading passages, charts and graphs, pictures, cartoons, and writing ability times are most easily adapted to this purpose.

2. Substantive items are those designed to directly measure knowledge about a population group. Such items might ask about the role of the Black church in Black life, the migration patterns of Chicanos in North America, or the factors that have led to the increasing numbers of women enrolling in graduate-level programs.

B. Evaluation Requirements

All test items will be reviewed and identifiable group reference items will be evaluated from the following perspectives:

1. Cognitive/Affective

These two dimensions should be considered when reviewing all group reference items. The cognitive dimension deals with the factual basis of item content and the affective dimension reflects the positive or negative feelings the item may evoke on the part of group members.

2. Controversial Material

Highly controversial issues, such as legalized abortion or hypotheses about genetic inferiority, must not be included in any test item unless they are both relevant and essential to effective measurement.

3. Examinee Perspective

All group reference items should be reviewed from the perspective of test takers who do not have access to an answer key. When an examinee must know the correct key to prevent an item from reinforcing negative attitudes, the item should be rejected. This situation most often arises when an item writer attempts to mislead test takers who hold stereotypical beliefs by using one or more distractors that are obvious stereotypes (obvious to the writer). In these cases, examinees who select a sterotype as the correct response are not routinely informed that their response was incorrect. Such practice may reinforce their belief in the legitimacy of the stereotype. This is particularly likely for examinees who subsequently receive a high score.

4. Stereotyping

All ETS tests will be reviewed to ensure that they do not contain language or symbols that reinforce sterotypes judged to be generally offensive. While it is clear that no ETS test intentionally contains blatant references to such offensive stereotypes, the potential exists for including material in a test that could be perceived as manifestations of such sterotypes by group members who take the test. Offensive stereotypes generally imply an inferiority or deficiency

of one of more groups in physical characteristics (for example, height, weight, attractiveness, strength) and psychological characteristics (for example, intelligence, ethics, emotions, behavioral patterns) generally regarded as desirable by the majority culture.

5. Caution Words and Phrases
   Through experience, those currently reviewing tests for sensitivity have learned that certain key words and phrases are more likely to accompany sensitive material. While for the most part use of these words and phrases in tests are proper and legitimate, they will receive extra attention by sensitivity reviewers when any two or more words are found in an item because they indicate an increased potential for including offensive material. Examples of such words are lower-class, discrimination, race, and housewife.

6. Special Review Criteria for Women's Concerns
   During the past decade, a great deal of progress has been made in identifying numerous manifestations of sexism in our society. This progress has included several efforts to identify and eliminate sexism in written language. Two notable efforts in this direction were the ETS Guidelines for Sex Fairness in Tests and Testing Programs and the McGraw-Hill Guidelines for Equal Treatment of the Sexes. Much of the material contained in these documents has been recast to serve as evaluation criteria. Sensitivity reviewers will ensure that all ETS tests are in compliance with these criteria.

7. Underlying Assumptions
   An underlying assumption is a subtle secondary premise in test material that reflects an individual's ethnocentric beliefs.

8. Context Considerations
   Reviews for sensitive material frequently require judgments relative to the context in which it is presented. In some cases, it may be necessary to measure one's knowledge of a domain by using material that some groups may feel is sensitive. There are four areas in which this occurs with some frequency: historical domain, literary domain, legal domain, and psychological domain.

Formal Structure and Procedures

The formal test sensitivity review is conducted by trained sensitivity reviewers who are often members of the Test Development staff at ETS. The review procedure involves an evaluation of the test by the sensitivity reviewer in accordance with the standard test evaluation criteria. Any comments and recommendations by the sensitivity reviewer are documented on the Test Sensitivity Review Form and returned to the test assembler. The test assembler discusses recommended changes with the sensitivity reviewer. If agreement is reached concerning changes, both sign and date the review form. If agreement cannot be reached, the matter is referred to the area Test Development Director and, if necessary, to an arbitration committee for resolution.

The sensitivity review of the CAT-ASVAB item pool followed the standard ETS procedures except for the final steps in which actual modifications are made to unacceptable test items. Rather, recommendations for changes by a minority panel were submitted for Government review in a Confidential Appendix to the ETS report, The Sensitivity Review of the CAT-ASVAB Item Bank (Wilson & Wainer, 1985.)

The CAT-ASVAB Sensitivity Review Board

The Sensitivity Review Board for this project was composed of twelve nationally renowned educators and test developers representing minority group concern. Asian American, Black American, Hispanic Americans, and both gender groups were represented on the panel.

The following ETS criteria were considered in selecting sensitivity reviewers:

A. Ability to perceive offensive material
B. Ability to review tests from multiple perspectives, not simply from the viewpoint of one group, or social/political perspective.
C. Coverage among the reviewers of key subject areas such as humanities and social sciences.

The Meeting

The CAT-ASVAB Sensitivity Review was conducted over a two-day period. The meeting began with a brief orientation session to familiarize panel members with the purpose, plan, and procedures for the sensitivity review. Each panel member had also received a booklet containing copies of relevant sections from ETS Test Sensitivity Review Process (Hunter & Slaughter, 1980) to review prior to the meeting. The ETS test evaluation criteria, guidelines for recognition of unacceptable sterotypes, caution words and phrases, and special review criteria for women's concerns were detailed in this booklet and were to be referred to by panel members as guidelines for item review and reporting of results.

Following the task orientation session, panel members were assigned to review CAT subtests in review teams. Each review team consisted of one male and one female of different ethnic affiliations, to ensure that the sensitivity review process was balanced across sex and ethnicity as much as possible. The CAT-ASVAB item bank consists of 2118 test items, representing nine different content areas: Electronics Information, Mechanical Comprehension, Shop Information, Automotive Information, Mathematics Knowledge, Arithmetic Reasoning, Paragraph Comprehension, Word Knowledge, and General Science. The 2118 times were divided up approximately equally among the reviewers, with each panel member reviewing between 334-363 items. Panel members were assigned to subtests in pairs so that each item would be reviewed by two people.

Description of Sensitivity Review Outcome

The following table was used to summarize the reviewers' conclusions:

	ACCEPTABLE		UNACCEPTABLE	
Subtest	As Is	With Revision	Potentially Offensive	Item Construction Flaw
Automotive Information				
Mathematics Knowledge				
Shop Information				
Mechanical Comprehension				
Electronics Information				
Arithmetic Reasoning				
General Science				
Word Knowledge				
Paragraph Comprehensive				
Totals				

Among the "acceptable" items some were judged to be acceptable only after modification. "Unacceptable" items were found to be so for two quite different reasons. Sometimes this was because of their potential offensiveness to some sub-groups of examinees. A second reason for finding an item unacceptable was because of a formal flaw in the item construction. This usually yielded the situation in which the number of correct responses to an item was unequal to one. Although some of the screened items had such flaws we left their description to another account; that aspect of the item pool is outside of the purview of this project. Therefore, "unacceptable" items in the CAT-ASVAB item bank, refer to items which were found to be unacceptable from a sensitivity review

perspective.

Due to the confidential nature of this project, the results of the sensitivity review as reported by panel members must be treated in a secure manner. Therefore, specific details on the test items and the sensitivity reviewers' comments and recommendations cannot be presented in this paper, but can be found in the Confidential Appendix to the Wilson & Wainer (1985) report.

However, the panel members did make some general comments and suggestions concerning the further improvement of the sensitivity review process itself.

In line with these and other suggestions we recommend in subsequent reviews the following additions/changes:

1) extend the review process sufficiently to allow a more extensive training period. This is important initially to fully inform the panel about the special aspects of CATs,
2) a slight modification of the review form to more closely correspond with the summary shown in the Table above.
3) have available to the review panel an information sheet on each item that contains:
   a) the item key,
   b) the content specifications that each item fulfills,
   c) the statistical results of the item's pretesting,
   d) comments of previous reviews (if any).

The availability of item history in a form as described in 3 above is pro forma in traditional test development. Although our recommendations deal with improving the ease and quality of the item review process, it has been the experience of test developers that such a procedure is quite useful throughout the item development process.

References

Hunter, R.V. & Slaugher C.D. (1980, July). ETS test sensitivity review process. Princeton, NJ: Educational Testing Service.

Wilson, S. & Wainer, H. (1985, September). The sensivitivity review of the CAT-ASVAB Item Bank. (Contract No. N66001-84-D-0083, D.O. 7J03). Submitted to Navy Personnel Research and Development Center. Princeton, NJ: Educational Testing Service.

MEASUREMENT PRECISION AND "RELIABILITY": SOME CONSIDERATIONS OF METRICS AND STOPPING RULES IN CAT

David Thissen
University of Kansas

Reliability has been defined as "the degree to which a test is free from error." While that definition is certainly attractive, it is not particularly mathematically tractable. In a more practical vein, classical test theory defined reliability as

$$rho^2_{XT} = rho_{XX'} = \frac{\text{"True score variance"}}{\text{"Observed score variance"}} \quad ; \qquad (1)$$

Since the error of measurement in classical theory is independent of true score, and has a constant variance, the form in equation (1) provides a standardized description of the relative size of the (lack of) error of the test score. It is also the "percentage of variance" in the observed scores which is due to true-score variation, and a predicted value of the correlation between observed scores on parallel forms. Here we will consider the extension of that concept to an IRT-scored Computerized Adaptive Test (CAT). This is not as hopeless as it may appear; CAT is sufficiently flexible that it can be made to produce results which meet the assumptions of classical test theory (which is something classical test scores never do).

The crucially wrong assumption of classical test theory, with respect to test scores, is that the error variance of all test scores is the same. The restriction of range of raw test scores alone makes that impossible, except in degenerate cases. However, in a CAT system, it is possible that all of the "test scores" could have the same error variance. Here we use the notation $\underline{est}(\theta)$ as a generic term for an estimate of θ; $\underline{est}(\theta)$ is some estimate of the location (like a mode or mean) of the likelihood over θ for a response vector, in which the likelihood is computed as the product of the trace lines for the responses. We use the notation $\underline{se}^2$ for the variance of that likelihood.

Equal $\underline{se}^2$ CAT

It would be difficult to construct a test giving constant $\underline{se}^2$s: the item pool would have to be so large and "broad" that, regardless of the ability of the examinee, the system could continue to administer increasingly easy or difficult items until the person gave sufficient numbers of both correct and incorrect responses to have a likelihood over θ with a constant variance. The CAT "stopping rule" would be a fixed value for the error variance for θ;

calibration of an item pool with the necessary properties is possible, but beyond the scope of this paper.

Under such a CAT system, strong parallels exist between the classical theory and the IRT CAT system, as summarized in Table 1.

TABLE 1

Classical Test Theory	IRT CAT (fixed se^2 stopping)
"True score"=T; var(T)	θ; var(θ)=1
"Error"=E; var(E)	error; var(error)=$se^2(\theta)$
"Observed score"=X; X = T + E, (independent), var(X)=var(T)+var(E).	est(θ); est(θ) = θ + error, (independent), var(est(θ))=1+$se^2(\theta)$.
rho^2_{XT} = var(T)/var(X) from (1).	rho^2 = 1/[1+$se^2(\theta)$] (2) from (1).

Using slightly different notation, Samejima (1977) notes that rho^2 in equation (2) represents the reliability, or expected correlation between parallel test scores, for an IRT-scored test. She notes that this form is impossible or deceptive if $se^2(\theta)$ varies as a function of θ; However, in a CAT system, it is possible that $se^2(\theta)$ is a constant and equation (2) is precisely correct.

Further parallels may be drawn between the classical theory and IRT CAT systems. In classical test theory, the best prediction of the true score is the so-called Kelley (1947) "regressed" estimate, which is

$$\text{rest}(T) = rho^2_{XT}\, X \qquad (3)$$

if X, T, and E all have mean zero and var(T)=1. If one "bends" IRT slightly, by claiming that the likelihood over θ is exactly (instead of approximately) Gaussian with mean est(θ) and variance $se^2(\theta)$, then

$$\text{rest}(\theta) = rho^2\ \text{est}(\theta) \qquad (4)$$

is exactly equal to either the Bayes modal estimate of θ or the expected a posteriori estimate of θ computed with a population distribution which is N(0,1). The mode and the mean become the same when "everything is Gaussian"; that is why we are using the generic notation est(θ) and rest(θ) for the un-regressed and regressed estimates of θ respectively.

Note that in the context of an IRT CAT system using a fixed (equal) $se^2(\theta)$ stopping rule, the concept of reliability is not "dead". It is, as a matter of fact, enhanced: it is "righter" than it ever was under the

classical theory. It is "righter" because it is based on the idea of equal error variance, which was never true for classical test scores, but which can be made true in CAT. The value of rho is both a prediction of the (hypothetical, "washed-brain") test-retest correlation between estimates of θ, and it is the "regression" or "shrinkage" constant in the Bayes estimates of θ.

Lord and Novick (1968) distinguish among three different "errors" in their discussion of the classical theory of reliability and those distinctions are useful here as well. The three error variances discussed by Lord and Novick, and their IRT counterparts in equal-se^2 CAT systems are tabulated in Table 2.

TABLE 2

Variance of Measurement

Classical Test Theory	IRT
$var(E) = var(X)[1-rho_{XX'}]$	$se^2(\theta) = var[est(\theta)][1-rho^2]$

Variance of Estimation

Classical Test Theory	IRT
$var(T)[1-rho_{XX'}]$	$1/[1+1/se^2(\theta)] = var[\theta][1-rho^2]$

Variance of Prediction

Classical Test Theory	IRT
$var(X)[1-rho^2_{XX'}]$	$1/[1+1/se^2(\theta)] + se^2(\theta) = var[est(\theta)][1-(rho^2)2]$

The discussion in Lord and Novick (1968, p. 67-8) is cast in terms of standard errors, instead of the variances given above; the standard errors are the square roots of these variances. The standard error of estimation is the square root of se^2; the standard deviation of the Bayesian posterior over θ is the square root of the variance of estimation, and the standard error of prediction of a subsequent estimate of θ from the regressed estimate is the square root of the variance of prediction. Those three values are not generally the same, which is one of the things wrong with defining reliability as "the degree to which a test is free from error": which error?. But they are all functions of rho^2.

It is also possible and reasonable in this context to consider the reliability of composite scores C obtained as

linear combinations of estimates of two or more θs. The simplest example is a composite of two tests. If

$$\underline{est}(C) = \underline{est}(\theta_1) + \underline{est}(\theta_2)$$

$$= (\theta_1 + error_1) + (\theta_2 + error_2),$$

and the θs are correlated r with each other and the errors are uncorrelated with everything, then

$$\underline{var}(\underline{est}(C)) = \underline{var}(\theta_1) + \underline{var}(\theta_2) + \underline{var}(E_1) + \underline{var}(E_2) + 2r\ \underline{sqrt}[\underline{var}(\theta_1)\underline{var}(\theta_2)]$$

and

$$\underline{var}\ (C) = \underline{var}(\theta_1) + \underline{var}(\theta_2) + 2r\ \underline{sqrt}[\underline{var}(\theta_1)\underline{var}(\theta_2)],$$

so

$$rho_{CC'} = \underline{var}(C)/\underline{var}(\underline{est}(C))$$

as per equation (1). Generalization to composites of more than two scales is obvious.

Other metrics

If the test scores are to be reported in some metric other than the θ-metric, such as "expected raw score" (E[score]), and composites are to be computed as linear combinations of scores in that metric, then it is obligatory that reliability be reported in the transformed metric. The simplicity described above can still be obtained in a CAT system if the stopping rule is based on equal $\underline{se}^2$s in the E[score] metric. This is somewhat deceptive psychologically: equal $\underline{se}^2$E[score] represents very unequal $\underline{se}^2$s in the θ metric, as E[score]s at the extremes have very small variances when the associated estimates on the ability dimension still have very large variances. But rho^2 would have its correct meaning in the metric in which the test is being described.

Unequal $\underline{se}^2$ stopping rules

CAT systems may be implemented with stopping rules other and those which give constant $\underline{se}^2$s. Such stopping rules clearly produce unequal $\underline{se}^2$s for different values of $\underline{est}(\theta)$. What happens to the concept, and the computation, or rho^2 under such alternative stopping rules?

There are several problems. The most obvious is that rho^2 cannot be computed as in equation (2), since that requires a constant value for $\underline{se}^2(\theta)$, which does not exist. Samejima (1977) suggests that for some purposes equation (2) may be replaced by

$$rho^2 = 1/[1+average(\underline{se}^2(\theta))] \qquad (5)$$

in which the error variances are averaged over the distribution of θ for the group being tested. Several further problems immediately arise. One is that the estimate of "reliability" obtained with equation (5) depends on the distribution of θ used in the "average"; if it is the theoretical population distribution, it depends on the theory, and if it is an empirical distribution, it depends on the sample.

A second problem is that it is not clear what purpose rho^2 computed using equation (5) could serve. It remains true that it is an estimate of the correlation that would be obtained between parallel tests. However, it is not an indicator of the size of the error variance for any particular $\underline{est}(\theta)$. Here, all $\underline{est}(\theta)$s have different $\underline{se}^2$s and rho^2 computed from (5) reflects only their average. An average of a set of numbers which are known to vary syatematically is not particularly useful. It would be much more useful, if the goal was to characterize the size of the errors of measurement, to abandon the concept of reliability altogether and report the sizes of the $\underline{se}^2$s as a function of θ, in either graphical or tabular form.

We noted above that, in the equal-$\underline{se}^2$ situation, the value of either $\underline{se}^2$ or rho^2 is indicative of the amount of "shrinkage" induced by the population distribution in regressed estimates of θ. Unfortunately, rho^2 from (5) is not informative about shrinkage. Indeed, there is rather serious theoretical problem underlying this loss-of-usefulness of the concept of reliability. When the $\underline{se}^2$s of $\underline{est}(\theta)$s are unequal, each is regressed a different amount (proportional to its own variance) in Bayesian estimation schemes. That means that the population distribution has differential effects on different individuals, which may be unfair.

Consideration of the reliability of composite scores is complicated extraordinarily by unequal $\underline{se}^2$s at the individual score level. Even in the simplest case in which a composite is the sum of two component scores, if those two component scores may have different $\underline{se}^2$s associated with different values, then a single value of the composite could have a wide variety of $\underline{se}^2$s. This is true because a single value of the composite could be produced by different combinations of θ_1 and θ_2, in which each θ could have different $\underline{se}^2$s which are combined to give the $\underline{se}^2$ for that particular way of obtaining that score on the composite.

While it is true that the distribution of $\underline{se}^2$s possible for each value of the composite could be averaged to give a single $average(\underline{se}^2)$ for that value of the composite score, that could be a fairly deceptive value when applied to any particular instance of that composite score. And each value of the composite score would still have a different $\underline{se}^2$; those would have to be averaged again to produce a single

"reliability coefficient" for the composite score as a whole. It is possible to do this numerically, or actually compute empirical test-retest reliability coefficients; but what is the use of such multiply-averaged numbers when we know that the average value applies to no particular test score?

Conclusion

There is a great deal to be said for the concept of reliability: it predicts test-retest correlation, describes the error variance of test scores, and specifies the amount of "shrinkage" inherent in Bayes estimators. There is even more to be said for the construction of tests which unambiguously have some specific reliability. It appears that such tests must be administered by CAT systems with a constant se^2 stopping rule.

References

Kelley, T.L. (1947). Fundamentals of statistics. Cambridge: Harvard University Press.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Samejima, F. (1977). A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

Multi-dimensional Performance Measurement

Lawrence M. Hanser and Jane M. Arabian
U.S. Army Research Institute[1]

Lauress Wise
American Institutes for Research

Introduction and Background

This paper is based on data collected for the large Army personnel research project titled "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Project A" (Eaton, Hanser, & Shields, 1985). This project was conceptualized and planned during the 1980 to 1981 time period, and a contract was signed with the prime contractor, Human Resources Research Organization (HumRRO), in 1982. It is being conducted jointly by scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), HumRRO, the American Institutes for Research (AIR), and Personnel Decisions Research Institute (PDRI).

Early in the planning for Project A, it was recognized that a large proportion of the research would have to be devoted to criterion development. Plans called for the development of several different measures of performance: (a) tests of hands-on performance, (b) paper and pencil tests of job knowledge, and (c) ratings of typical performance. Each of these broad categories of criteria were further subdivided. Hands-on tests included tasks which were specific to each Military Occupational Specialty (MOS) as well as tasks common to all MOS. Two kinds of paper and pencil tests were constructed: (a) to emphasize the content of formal school training, and (b) to emphasize MOS-specific task performance. Rating forms were constructed both for MOS-specific task performance as well as for non MOS-specific Army-wide performance that we have labelled broadly as "soldiering."

The initial impetus for developing such a comprehensive set of criterion measures was largely a function of our underlying theory of performance measurement. This underlying theory states rather simply that performance in a job is multi-dimensional, and that it is not possible to capture that multi-dimensionality using only one measurement method. A method of measurement may be intrinsic to some tasks. For example, having the requisite knowledge of how to take a person's blood pressure may not be the same as actually being able to perform the task accurately, yet both are important. An individual may score high on a paper and pencil test on this task, but might not score as high on a hands-on test of this task. In order to be successful in performing this task on the job it requires: (a) the knowledge of how to do the task, (b) the physical skills to perform the task, and (c) the motivation to do it. Or to put it in another well known way: performance = f(ability x motivation).

---

[1]The views expressed in this paper are those of the author and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army

Because of the complexity of the criterion space being measured in this project, it is extremely important that it be fully understood prior to choosing a final set of predictors and recommending changes to the Army's selection and classification procedures. Several recent papers by project scientists have begun to address the issues associated with criterion development (c.f., Borman, White, Gast, & Pulakos, 1985; Campbell & Harris, 1985; Rumsey, Osborn, & Ford, 1985). Borman et al. constructed and tested a path model of supervisory and peer ratings to examine how each are related to other measures of performance. They found that both job knowledge and hands-on task proficiency are related to ratings, with the dominant path between ratings and hands-on proficiency. They conclude, however, that "... for the most part different methods of measuring job performance yield quite different results." Campbell and Harris describe the results of attempting to interpret criteria using a group of "concerned psychologists." They also present a "working model of job performance for the domain of skilled jobs." In examining the correlation matrices of hands-on and job knowledge tests and rating scales, they state "... the methods correlate more highly within themselves than they do across measures." Rumsey et al. examine the relationships between job knowledge tests and hands-on tests of job proficiency. In each of these papers, a central theme is the multi-dimensionality of performance and the importance of using different measurement methods to capture performance adequately.

The intent of this paper is to further explore the criterion space measured in Project A. Previous research has focused on aggregate measures of performance such as total scores on hands-on or paper and pencil tests or average ratings across several dimensions. In this paper we focus on task level measures in order to begin to understand better the relationships between kinds of tasks and methods of measuring performance on them. Through this we hope to gain a better understanding of the method variance associated with measures of task performance.

Method

Subjects

Data reported in this paper were collected in 1984 as part of field tests of the criterion measures developed by Project A scientists. Participants included first tour soldiers in two Army MOS: (a) 178 Infantrymen (MOS 11B) and (b) 167 Medical Specialists (MOS 91A). A complete description of the data collection methods can be found in Campbell and Harris (1985).

Variables

Percent correct steps per task and average supervisory rating per task provided the major variables used in these analyses. Percent correct scores were obtained on both hands-on and written tests. For each MOS reported here, approximately 15 tasks were scored using all three measurement methods: (a) hands-on performance, (b) multiple choice paper and pencil test, and (c) average supervisory rating of task performance. Approximately 15 additional tasks per MOS were tested in the paper and pencil test, and these were also included in the analyses. In addition, total score on a paper and pencil test focusing on training course content, average supervisory rating on overall performance, and Armed Services Vocational Aptitude Battery (ASVAB) subtest standard scores were included. This resulted in a total of approximately 71 variables per MOS

to be included in these analyses. Although these are a relatively small number of subjects given the number of variables, the limits of analysis are a function of the number of factors extracted. These sample sizes will support the extraction of a maximum of five to seven factors per MOS.

Analyses

Though some "feel anxious in the presence of too many partial or semi-partial correlations" (Campbell & Harris, 1985), we decided to explore these data using factor analysis. Our specific plans were as follows: (a) extract a set of oblique factors for each MOS, (b) examine the inter-factor correlation matrices, and (c) examine the patterns of loadings within and across MOS. We used a principal axis solution with an iterative solution for the communalities and a Promax rotation. We decided on the number of factors to extract based on an inspection of the scree and interpretability of various solutions. In order to conserve space, descriptive statistics and reliabilities are not reported here. They are, however, available elsewhere (Borman et al., 1985; Campbell & Harris, 1985; Rumsey et al., 1985).

Results and Discussion

The data on the Medical Specialists yielded a five factor solution. Table 1 shows the oblique solution. Variables reported in the table are limited to the three highest loading on any factor, any variable with an absolute loading of greater than .30 on a cross-method factor, and any variable with loadings greater than .30 on two or more factors.

Table 1. Rotated Factor Pattern (STD REG COEFS)

I	II	III	IV	V	
80	.	.	.	.	Rating:Splint Suspected Fracture <Supv>
77	.	.	.	.	Rating:Put on Field/Pres Dressing <Supv>
75	.	.	.	.	Rating:Perform CPR <Supv>
58	.	.	-35	.	Rating:Measure/Record Respir. <Supv>
53	.	30	.	.	Rating:Measure/Record Pulse <Supv>
.	57	.	.	.	P&P:D9-Replace Filters in M17 Mask
.	51	.	.	.	P&P:I4-Measure/Record Respirations
.	47	.	.	.	P&P:I9-Estab/Maintain a sterile fld
.	43	.	.	.	HO: A4-Put on Field/Pres Dressing
.	34	.	.	.	HO: A9-Init a Field Med Card
.	.	68	.	.	ASVAB SUBTEST SCR-Arithmetic Reasoning
.	.	57	.	.	ASVAB SUBTEST SCR-Math Knowledge
.	.	52	.	.	ASVAB SUBTEST SCR-Coding Speed
.	.	49	.	.	P&P: I6-Assemble Needle & Syringe
.	.	49	.	.	P&P: K2-Draft/Fire TPR Charts
.	.	42	.	.	P&P: A6-Open Airway
.	.	40	.	.	P&P: I7-Change a Sterile Dressing
.	.	41	32	.	School: All Items
.	.	.	76	.	ASVAB SUBTEST SCR-Auto/Shop
.	.	.	71	.	ASVAB SUBTEST SCR-Electronics Information
.	.	.	59	.	ASVAB SUBTEST SCR-Mechanical Comprehension
.	.	.	37	.	P&P: G3-Vehicle Recognition
.	.	.	.	68	HO: I3-Measure/Record Pulse
.	.	.	.	51	HO: I9-Est/Maintain Sterile Field

.	.	.	.	47			HO: I4-Measure/Record Respir.
.	.	.	33	35			HO: AB-Splint Suspected Fracture

As expected, there are strong method factors, with little overlap of wariables across method factors. Note, however, that two ratings overlap with the ASVAB factors, and one of the hands-on tasks overlaps with an ASVAB factor. Two hands-on tasks have loadings greater than .30 on Factor II, the paper and pencil job knowledge test factor. Several of the job knowledge test tasks load on the two ASVAB factors. Also, ASVAB splits into two factors, a math/speed factor and a technical factor. Table 2 provides the factor correlations.

Table 2. Inter-Factor correlations

	I	II	III	IV	V
I	100	1	7	-11	17
II	1	100	15	27	-2
III	7	15	100	-6	19
IV	-11	27	-6	100	-8
V	17	-2	19	-8	100

Not surprisingly, the paper and pencil job knowledge test factor, Factor II, and an ASVAB factor, Factor IV, have the highest correlation. Note, however, that none of the ASVAB subtests have loadings of .30 or higher on Factor II, and that it is the ASVAB technical factor which correlates highest with the job knowledge paper and pencil test factor. The ASVAB Verbal subtest did not meet the criteria for inclusion in this table. These results would seem to indicate that correlations between ASVAB and paper and pencil job knowledge measures are not simply the result of shared method variance.

The next highest inter-factor correlations are between the hands-on factor, Factor V, and the ASVAB math/speed and rating factors, Factors III and I respectively. While the hands-on factor is a relatively pure method factor, its correlations with the other factors strengthen the conclusions of Borman et al. Each method appears to measure a different but related piece of job performance.

Table 3 contains the oblique promax factor pattern for Infantrymen. Seven factors were extracted. The choice of variables to report was based on the same rules as for the previous table of loadings.

Table 3. Rotated Factor Pattern (STD REG COEFS)

I	II	III	IV	V	VI	VII	
64	.	.	.	.	.	.	P&P: E5-Oper as Station in Radio Net
64	.	.	.	.	.	.	School: All Items
61	.	.	.	.	.	.	P&P: B4-Perform OP Maint. on M16A1
59	.	.	.	.	.	.	P&P: H1-Perform Tracked Vehicle Maint
56	.	.	-39	.	.	.	P&P: E1-Collect/Report Info
.	66	.	.	.	.	.	Rating: Install/Fire/Recover M18A1 <Supv>
.	65	.	.	.	.	.	Rating: Load/Clear M60 <Supv>
.	59	.	.	.	.	.	Rating: Prepare Range Card for M60 <Supv>
.	54	.	37	.	.	.	Rating: Mean non MOS-Specific<Supv>
.	50	.	33	.	.	.	Rating: Navigate on Ground <Supv>
.	38	.	39	.	.	.	Rating: Set Headspace/Timing on .50 <Supv

.	31	.	.	.	.	.	P&P: G8-Estimate Range
.	.	76	.	.	.	.	ASVAB SUBTEST SCR-Auto/Shop
.	.	74	.	.	.	.	ASVAB SUBTEST SCR-Mechanical Comprehension
.	.	73	.	.	.	.	ASVAB SUBTEST SCR-General Science
.	.	73	.	.	.	.	ASVAB SUBTEST SCR-Verbal
.	.	.	79	.	.	.	Rating: Op as Station in Radio Net <Supv>
.	.	.	76	.	.	.	Rating: Op Radio Set AN/PRC-77 <Supv>
.	.	.	44	.	.	.	HO: E5-Op as Station in Radio Net
.	.	.	31	.	.	.	HO: BC-Engage Targets w LAW
.	.	.	.	68	.	.	HO: C6-Call/Adjust Indirect Fire
.	.	.	.	67	.	.	HO: G8-Estimate Range
.	.	.	.	55	.	.	HO: B4-Perform Op Maint on M16A1
.	39	.	.	37	.	.	Rating: Call/Adjust Indirect Fire <Supv>
30	.	.	.	32	.	.	P&P: B9-Engage w Hand Grenades
32	.	.	.	.	.	.	HO: BB-Prepare Range Card for M60
.	.	.	.	.	58	.	HO: J1-Movement in Urban Terrain
.	.	.	.	.	56	.	HO: BA-Prepare Dragon for Firing
.	.	.	.	36	50	.	HO: B9-Engage Targets w Grenades
.	.	.	.	.	47	.	HO: I1-Install/Fire/Recover M18A1
.	.	.	.	.	35	.	P&P: BA-Prepare Dragon for Firing
.	.	.	.	.	.	71	ASVAB SUBTEST SCR-Numerical Operations
.	.	.	.	.	.	59	ASVAB SUBTEST SCR-Coding Speed
.	.	40	.	.	.	54	ASVAB SUBTEST SCR-Math Knowledge
.	.	41	.	.	.	53	ASVAB SUBTEST SCR-Arithmetic Reasoning

While similar method factors emerge, the factor space for infantrymen is slightly more complex. The ASVAB Factors III and VII are quite clean, though Factor III and the paper and pencil job knowledge test Factor I are relatively oblique (Table 4.). These factors are substantially more correlated than are the two ASVAB factors with each other. Note also that the ASVAB math/speed Factor VII has a lower correlation with the paper and pencil job knowledge test Factor I, than the more technical ASVAB Factor III. If there is a simple "written test" factor, it failed to emerge in either of these solutions.

Perhaps most interesting are Factors IV and V. Each of these factors has a mixture of variable loadings representing different measurement methods. On Factor IV the supervisory rating and hands-on test for operating as a radio station in a net both load substantially. On Factor V the supervisory rating and hands-on test for call/adjust indirect fire both load substantially, and the paper and pencil and hands-on tests for engage targets with grenades also both load substantially.

Table 4 gives the correlations among the factors for Infantrymen. This solution is considerably more oblique than the solution for Medical Specialists.

Table 4. Inter-Factor correlations

	I	II	III	IV	V	VI	VII
I	100	30	53	36	18	25	24
II	30	100	13	40	18	19	-3
III	53	13	100	6	13	-1	29
IV	36	40	6	100	21	34	-5
V	18	18	13	21	100	-2	1

VI	25	19	-1	34	-2	100	0
VII	24	-3	29	-5	1	0	100

The highest correlation is between Factor I, the paper and pencil job knowledge test factor, and Factor III, the ASVAB technical factor. This result is similar to that noted previously. The two primarily supervisory rating factors, II and IV, are quite highly correlated with the paper and pencil test of job knowledge factor. In fact, Factor IV correlates almost as highly with Factor I ($r=.36$) as it does with the other rating factor, Factor II ($r=.40$). The two hands-on test factors, V and VI, are uncorrelated with each other. Factor VI has respectable correlations with both the paper and pencil job knowledge test factor, Factor I, and the rating factor, Factor IV.

Conclusions

Our tendency as psychologists is to abhor method variance as something to be avoided. This should not necessarily be the case in the realm of job performance measurement. Performance of a task requires first the ability and motivation to learn the task, and second the skill, ability, and motivation to perform it. Different methods of measuring performance, hands-on tests, written tests, and ratings, capture slightly different aspects of performance. Some of these relationships are apparent from the data presented above.

What remains for us is to understand which kinds of tasks are most appropriately measured by which methods. The research reported here, while open to several interpretations, presents a method and several examples of a way to do this. Clearly, more research needs to be conducted into the content of the tasks themselves and their relationships to method factors across several more occupations than are included here.

References

Borman, W. C., White, L. A., Gast, I. F., & Pulakos, E. D. (1985, August). Performance Ratings as Criteria: What is being Measured. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Campbell, J. P., & Harris, J. H. (1985, August). Criterion Reduction and Combination via a Participative Decision Making Panel. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Eaton, N. K., Hanser, L. M., & Shields, J. L. (in press). Validating Selection Tests Against Job Performance. In J. Zeidner (ED.), Human Productivity Enhancement. New York: Praeger.

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985, August). Comparing Work Sample and Job Knowledge Measures. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

WALK-THROUGH PERFORMANCE TEST DEVELOPMENT: LESSONS LEARNED

Carl J. Taylor
Jack L. Blackhurst
Rodger D. Ballentine

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

The Air Force Human Resources Laboratory (AFHRL) is involved in a multi-year effort investigating the feasibility of measuring and linking job performance to Armed Services Vocational Aptitude Battery (ASVAB) scores. The major focus of this work is the development of a technology for systematically obtaining job performance data as criteria for validating enlisted, officer, and civilian selection systems, and evaluating Air Force training programs.

Planning for the Air Force's program of research in performance assessment began several years ago as the result of three primary requirements. Operational military and civilian program managers in the manpower, personnel, and training communities asked AFHRL to develop an approach for measuring job performance so that the measures could be used to assist in the evaluation of their training and selection programs. Secondly, the manpower, personnel, and training research community needed performance measures to serve as criteria in their research projects. Plans for the Air Force performance measurement effort to meet these requirements were already under development when a third requirement for these measures came with the congressional mandate to test the feasibility of validating the ASVAB against job performance measures.

The cornerstone of this criterion development effort is a work sample testing approach known as Walk-Through Performance Testing (WTPT). The WTPT process is being developed to expand the range of job tasks measured to include tasks which do not lend themselves to hands-on testing because of cost, time, and/or safety considerations. WTPT combines hands-on task performance and interview procedures to provide a high fidelity measure of individual technical job competence.

The walk-through procedure involves taking the job incumbent to the work site and administering a combination of performance tasks and interview questions. The interview testing component will be evaluated both as supplement to hands-on data collection to ensure adequate sampling of the domain of tasks in a job and as a more cost-effective substitute for hands-on testing. A wide range of alternative job performance measures will be developed in addition to the walk-through testing methodology. These include peer, supervisor, and self-performance ratings, at the task, dimension, and global levels.

Alternative measures will be developed for the same specialties used to develop walk-through testing techniques. When job sample, interview, and alternative forms have been developed for each of the Air Force specialties selected for this study, their relative utility will be determined. Existing performance measures such as technical training scores, Airman Performance Report ratings, skill-level advancement indices, and Specialty Knowledge Test scores will also be evaluated as possible alternative measures.

Over the past two years, AFHRL has developed WTPTs for four Air Force Specialties (AFS). The Jet Engine Mechanic career field was selected as the AFS for WTPT prototype development. As the jet engine mechanic WTPT was developed, the process was documented (Alba and Wilcox, 1985) to serve as a set of procedural guidelines. These guidelines were then applied to three additional career fields: Air Traffic Control Operator, Avionic Communications Specialist, and Ground Radio Operator. The desire was to apply the guidelines directly to the next three career fields; however, it was anticipated that modifications might be needed to account for the dissimilarities between mechanical and non-mechanical career fields. The purpose of this paper is to highlight the lessons learned as the WTPT Procedural Guidelines were applied to the other career fields.

TASK SELECTION/TASK ANALYSIS

This step of WTPT development was directly transferable to the three additional career fields. That is, computerized occupational analysis data was used to develop a task selection plan. This plan specified the guidelines for identifying Phase I tasks (tasks performed by 30% or more of first-term incumbents) and Phase II tasks (tasks performed by 40% or more of first-term incumbents in a functional area and not included in Phase I).

The selected tasks were then reviewed/validated by subject matter experts (SME) during task selection workshops. These workshops resulted in a number of tasks being discarded or moved from one phase to another due to mission requirement changes, technology changes, or low task difficulty levels. The inconsistencies between occupational analysis data and SME input can be attributed to the age of the data. Even though the occupational survey reports (OSR) were the most recent ones for the three career fields, they were three to four years old. This resulted in several false starts. For instance, based on OSR data pertaining to the numbers of first-term personnel in subareas of the ground radio operator career field, the initial focus of task selection was on the Military Affiliated Radio System (MARS) area and not on the Mobile Communications (MOB) area. After spending some time with SMEs, it was discovered that, since the last OSR, mission emphasis had been increasing for the MOBs and decreasing for the MARS.

This type of problem can best be alleviated in the future by ensuring that OSR data is current on career fields selected for WTPT development. Communication with the major command (MAJCOM) career field functional manager very early in the task selection plan development is also essential so that recent or upcoming modifications in technology or mission can be a built into the WTPT.

Once tasks had been selected, a task selection workshop was held. Participants included the career field functional manager and SMEs having extensive experience in the career field. During the task selection workshop, emphasis was placed on ensuring the tasks were currently performed and that they were performed in a similar manner by everyone in the applicable functional area (i.e., flightline, shop, operations, maintenance). If workshop participants suggest deleting a task or moving it from one phase to another, detailed documentation should be provided on the justification. For each task, documentation should also be provided on equipment-related similarities/differences in how a task is performed, reliance on local operating procedures and Air Force or MAJCOM regulations, whether

first-termers only perform parts of or all of a task, and if a task is performed in the same manner in all units. Increasing the emphasis in these areas during the workshop will reduce confusion during the task analysis stage.

The objective of our task analysis is to gather information essential to WTPT item development. Relevant information includes beginning and ending points for each specific task, critical steps for task accomplishment, logistical requirements for task completion, required configuration of equipment, time critical and safety steps, effects of local operating procedures on task performance, and representativeness of the task. This information is gathered by referencing applicable regulations, technical orders, and local operating instructions as well as discussions with SMEs in the field. The desired result of the analysis is a comprehensive list of steps required for successful task completion. These steps should be generalizable to any situation in which the task might be observed and should be used to objectively evaluate an individual's performance on the task.

For the ground radio and avionic communications career fields this process worked very well; however, problems were encountered in the air traffic control area. One problem area is related to the differences in the work environment for radar approach control (RAPCON) personnel and tower personnel. RAPCON controllers depend solely on various radar scopes to separate and sequence aircraft. Tower controllers are concerned with line of sight traffic separation and use radar very infrequently. The result is that there are only a very small number of Phase I tasks (i.e., those accomplished in both facilities). Those tasks which are performed in both areas, typically are performed differently in the two areas. This resulted in a "separate but equal" approach. Problems encountered in constructing the actual WTPT will be discussed in the next section.

Test Development

The WTPT procedural guidelines also served as the basis for test development on all three additional specialties. The guidelines transferred easily to the new specialties; however, problems were encountered in developing standardized tests once the tasks were identified.

The first problem area is test security. Because of the physical arrangement of most radio operations rooms, avionic maintenance shops, and air traffic control facilities, it will be difficult to administer the WTPT without allowing others to observe the tasks being evaluated. One possible solution is to temporarily partition off the test area; however, this may impact the unit's ability to carry out their mission. Additionally, some radio facilities require security clearances for facility access due to their mission. This may present problems during data collection as prospective contractor evaluators have not obtained clearances. In an attempt to alleviate this problem, the AFHRL recently conducted a study comparing performance ratings given by contractors with those given by active duty personnel; however, results are unavailable at this time.

The air traffic control field presents unique problems in terms of developing standardized testing situations. For instance, the radio, avionics, and jet engine mechanic jobs involve heavy reliance on technical orders, are labor intensive, and are characterized by completion of tasks in a routine, standardized manner. Most tasks can only be done one way and only

rarely does the performance of one task have any effect on subsequent tasks. These characteristics make it relatively easy to construct standardized testing environments. In addition, jet engines or radio equipment can actually be used for the test and the test administrator can ensure the equipment is configured exactly the same for each incumbent.

Conversely, the air traffic controller's job is characterized by many correct ways to deal with a situation, is intensely interactive, and often has a time critical component. For test administration purposes it cannot be requested that aircraft fly repeatedly in preset patterns for purposes of presenting standardized scenarios to the incumbents being evaluated.

Another characteristic of the air traffic control area is that individual tasks for this career field require few measureable steps for completion. This is a function of the air traffic controller's job. For instance, tuning a radio or troubleshooting avionics equipment requires a series of specific steps which can be readily observed and objectively evaluated. In contrast, the observable portions of the air traffic control job consists of keying a microphone, operating runway lights, or using the correct phraseology. These tasks cannot be further dissected into measureable substeps.

To address these problems, AFHRL developed a "job module approach" for air traffic control operators involved in RAPCON duties. The job module makes use of RAPCON simulators and combines a number of tasks into standardized scenarios. The incumbent is seated at the simulator radar screen and is required to "work" typical traffic problems of varying levels of complexity. Individual tasks are scored as they occur in these scenarios. Due to the absence of tower simulators, a number of approaches for developing standardized job modules for tower personnel were considered. These approaches will be outlined in the following section.

Due to the absence of tower simulators, AFHRL considered a number of approaches to standardized hands-on testing for tower personnel (e.g., using the tower simulator, video taping scenarios from the simulator, developing a new simulator, computer games, slide presentations, and video taping live situations). Each of these approaches was evaluated with respect to test standardization, objectivity, stimulus fidelity, cost, discrimination, and time to develop. As a result of these comparisons, the use of video taped live traffic is being pursued.

Summary

The AFHRL has developed a set of procedural guidelines which can be used to design valid, reliable, and standardized hands-on performance assessment instruments for Air Force enlisted career fields. The guidelines were developed around a mechanical career field and then applied to career fields representing the remaining three ASVAB Aptitude Index areas (administrative, electrical, and general). The major lesson learned with respect to applying the procedural guidelines concerns the OSR information used as the starting point for task selection. Every effort should be made to ensure that the OSRs are current for future AFSs. The other lessons learned involve the ability to develop standardized testing environments and test administration logistics issues. Being aware of these issues at the beginning of WTPT development will result in a much smoother development.

References

Alba, P. A. & Wilcox, T. (1985). Walk-through performance testing procedural guidelines manual. Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Uniform Guidelines for Employment Selection Procedures. From Federal Register 43(166), August 15, 1978.

ON THE CONTENT AND MEASUREMENT VALIDITY OF HANDS-ON JOB PERFORMANCE TESTS

PROBLEM

The justification for using aptitude tests to help select enlisted recruits and assign them to occupational specialties is that aptitude tests are valid predictors of performance. The aptitude tests used by the military services have been extensively validated as predictors of performance in occupational specialty training courses. Their usefulness as predictors of performance on the job, however, is less well documented. The Job Performance Measurement Project has been initiated to validate the Armed Services Vocational Aptitude Battery (ASVAB) as a predictor of job performance.

The question then arises of how job performance should be measured. The measures favored by the Joint Service Job Performance Measurement Working Group are hands-on job performance tests. These tests have intrinsic validity because of their high fidelity to job behavior. Hands-on performance tests, however, are susceptible to poor content and measurement validity.

- Poor content validity may arise because the tests focus on skills easy to test in the hands-on mode without including the full range of job requirements.

- Poor measurement validity may arise because the scoring standards of test administrators are not calibrated and because test security is difficult to maintain (examinees can find out what is being tested and practice beforehand).

The purpose of this analysis is to examine the content and measurement validity of prototype hands-on tests for three Marine Corps specialties – Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman – used in a feasibility study to evaluate ASVAB qualification standards.

FINDINGS

The findings pertain to the two technical specialties, Ground Radio Repair and Automotive Mechanic. Because the infantry riflemen in the sample had limited job experience, the results for them are inconclusive.

- Hands-on test scores were only weakly related to amount of job experience, as measured by months in the Marine Corps (figure I). Test scores were expected to increase with experience, and the lack of relationship raises questions about how well the tests represent the full range of job requirements.

- The ASVAB is a valid predictor of hands-on test scores for people with 2 years or less of service in the Marine Corps, but not for people with more than 2 years of service (table I).

- Hands-on test scores did increase with experience for people with low aptitude, but not for people with high aptitude (figure II).

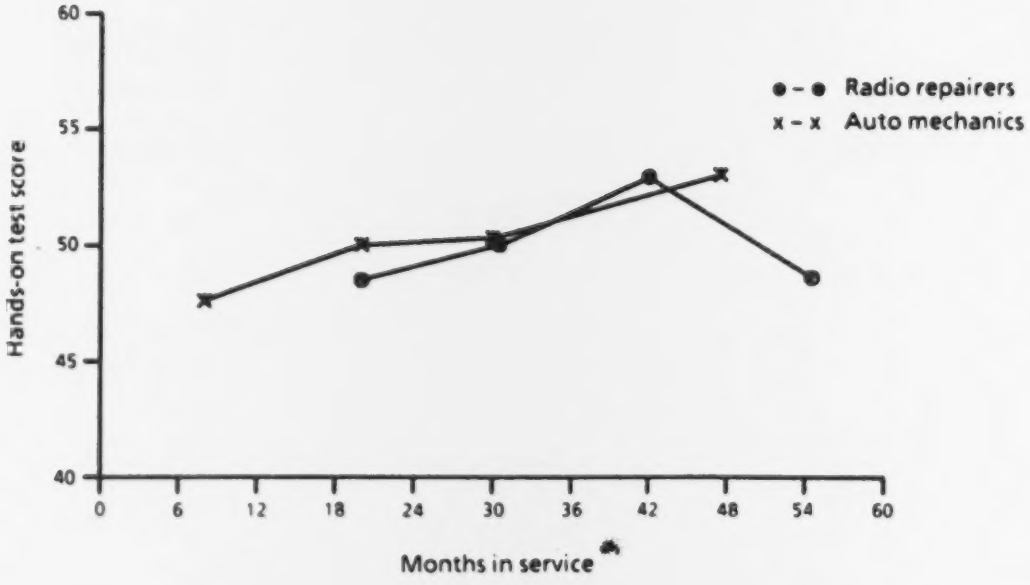


FIG. I: JOB PERFORMANCE RELATED TO TIME IN THE MARINE CORPS

These results suggest that the hands-on test content was appropriate for people recently assigned to their first duty station, but less appropriate for people with more experience, who perform job tasks not reflected in the tests.

The findings on measurement validity bear on institutionalizing of hands-on job performance tests:

- The test administrators used different scoring standards, and the same administrators changed their scoring standards across time (see figure III).
- Maintaining test security is difficult.

TABLE I

VALIDITY OF THE ASVAB FOR PREDICTING HANDS-ON TEST SCORES

Months in service	Validity[a]	Number of cases
	Ground Radio Repair	
15-25	.69	38
26-35	.00	53
36-48	.00	37
Total	.37	128
	Automotive Mechanic	
2-14	.72	57
15-25	.52	56
26-34	.15	53
35-60	-.07	54
Total	.37	220

a. Population-wide estimate of validity coefficient.

CONCLUSIONS

- The ASVAB is a valid predictor of job performance, as measured by hands-on tests.
- But hands-on tests lack robustness:

- Content validity is sensitive to job experience.
- Measurement validity is sensitive to the calibration, or scoring standards, of test administrators.

- Institutionalizing hands-on job performance tests would be difficult.

Milton H. Maier
Center for Naval Analyses
Alexandria, Virginia 22302-0268

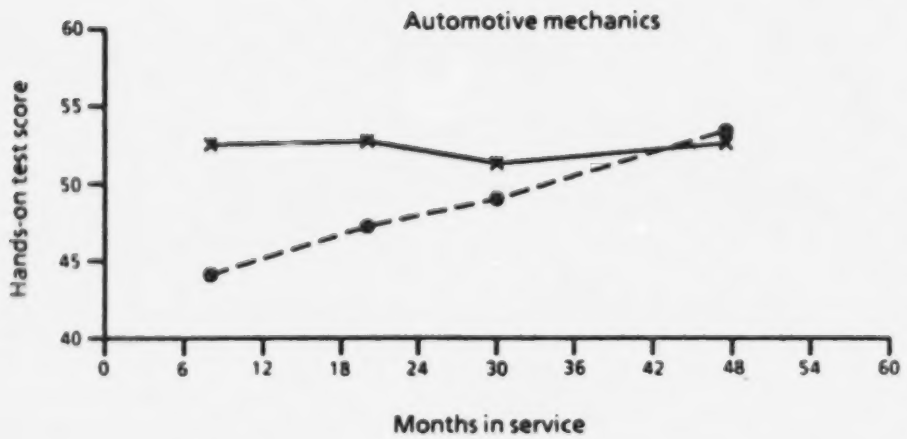


FIG. II: PERFORMANCE RELATED TO APTITUDE AND EXPERIENCE

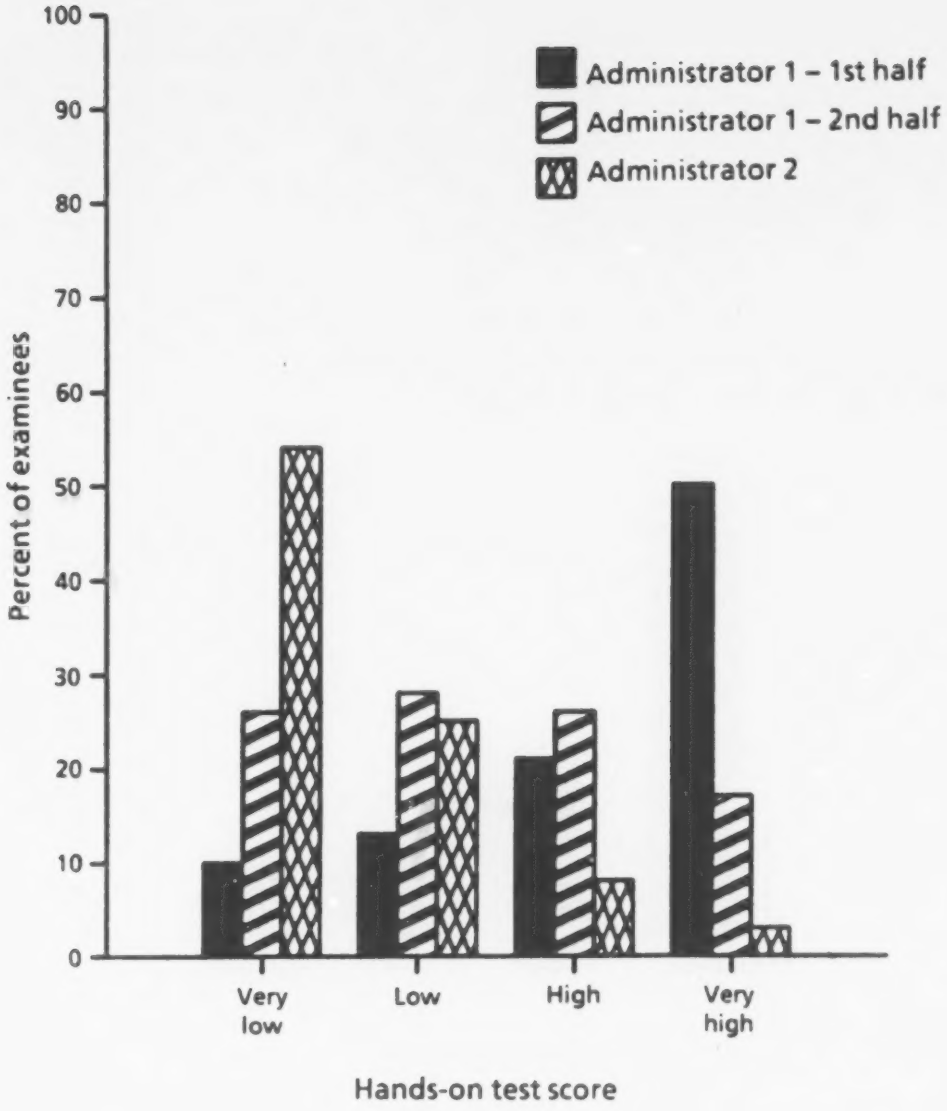


FIG. III: HANDS-ON TEST SCORES ASSIGNED BY TEST ADMINISTRATORS TO AUTOMOTIVE MECHANICS

DEVELOPING AND EVALUATING A HANDS-ON PERFORMANCE TEST

L. P. Kroeker, R. M. Bearden, & G. J. Laabs
Navy Personnel Research and Development Center
San Diego, CA 92152

The first step in reaching the Navy's goal to improve enlisted personnel classification through the use of job performance information includes the development and evaluation of hands-on performance tests. This paper describes the developmental procedures employed in the first rating to be covered in a Congressionally mandated, Joint-Service project aimed at linking job performance to enlistment standards.

Background

In 1982, Congress formally required the Armed Services to establish methods for measuring job performance and validating selection standards against them. The Navy's contribution to this co-ordinated effort is entitled Performance-Based Personnel Classification. It's objectives are to investigate measurement approaches that can be used to assess on-the-job performance and to improve the Navy's automated classification and assignment system or CLASP (Kroeker and Rafacz, 1983) by including job performance information.

The Navy's approach focuses on direct measurement of technical proficiency, which follows the research strategy of the Joint-Service project. The purpose of this large scale effort is to develop job performance measures for first-term enlistees with four (or fewer) years of service and demonstrate their use as criteria for predictor validation (Office of the Secretary of Defense, 1984).

In the Joint-Service project the hands-on job sample test has been adopted as a high fidelity benchmark measure against which other less costly and more easily administered measures will be compared. Therefore, it is essential to construct valid, reliable, and objective hands-on performance tests.

Because of the extensive resources required for this large scale validation effort, performance measures will be developed only for a small number of Navy ratings. The Machinist's Mate (MM) rating is the first one to be covered.

Machinist Mate Performance Test Development

To ensure that the job sample test for MM's adequately represented the technical proficiency content domain, a sequence of task selection steps were taken (Guion, 1979). These steps, which involved the orderly reduction of the job content universe to the job sample, resulted in the identification of the critical tasks to be included in the test.

The job content universe was defined by a comprehensive job task analysis taken from the Navy Occupational Task Analysis Program (NOTAP) data base. The latter was supplemented with information from: (1) Occupational Standards; (2) Personnel Advancement Requirements (PAR); (3) Personnel Qualification Standards (PQS); (4) standard operating procedures; (5) standard maintenance procedures; (6) technical manuals; and (7) A-school learning objectives.

The job content domain was determined by subject matter expert (SME) judgments. The original task list was reduced to only those tasks involving technical proficiency for MM's employed on 1052 class frigates.

The next step involved the definition of the test content universe. This universe consisted of those job tasks from the technical proficiency content domain that might be observed in a hands-on test, in addition to all conditions that might be imposed in the testing situation and the procedures for observing and recording responses. At this stage SME judgments were gathered on the criticalness of each task to the operation of a ship's propulsion plant.

Finally, the test content domain was defined. Tasks were drawn from the test content universe based on the task-critical judgments obtained in the preceding step. They cover the two main areas of work behavior for MM's, namely, maintenance and watch-standing. The procedure ultimately yielded the following tasks, which were approved as a representative job sample by the MM training community:

A. Maintenance Tasks

1. Planned Maintenance
   - (i) Sample oil
   - (ii) Inspect oil
   - (iii) Tag out

2. Corrective Maintenance
   - (i) Make gasket
   - (ii) Repair valve
   - (iii) Repack Valve

B. Watchstanding Tasks

1. Normal Watchstanding
   - (i) Shift and inspect water strainer
   - (ii) Shift and inspect oil strainer
   - (iii) Operate eductor
   - (iv) Operate fire pump
   - (v) Record gauge readings
   - (vi) Clean and inspect oil filter or operate oil pump

2. Casualty Control
   - (i) Loss of oil pressure
   - (ii) Major oil leak
   - (iii) Loss of vacuum
   - (iv) Hot bearing

Upon completing the task selection procedures the following developmental procedure was followed. First, the behavioral elements of each task were identified. This was relatively easy because MM's have fully proceduralized job performance aids that specify the step-by-step requirements of watchstanding tasks. These task steps are outlined in the Engineering Operational Procedures (EOP) and Engineering Operational Casualty Control (EOCC) documents. The specification of the behavioral elements for the maintenance tasks were obtained from fleet SME's and verified by the MM training community.

Following the delineation of the behavioral elements associated with each task, MM's aboard two frigates were observed as they performed the tasks. A few tasks were eliminated because of the lack of procedural standardization across ships. It was at this point that the set of tasks were presented to the MM training community for approval.

After the performance of the tasks was observed aboard the two ships, a structured observation form was generated. Draft versions were reviewed for accuracy and comformity to conventional operating procedures by SME's. Next an experimental version of the instrument was administered and revised.

The experimental version was administered to apprentice MM's on board a frigate. As a result of this try-out some modifications to the test were made. One of our discoveries was that there was informal job specialization even on this relatively small ship, whereas we had expected all first-term MM's to perform almost all of the watchstanding tasks connected with the MM propulsion engineering spaces. This led to the development of parallel forms for the two major MM work spaces: one for the Engine Room and one for the Auxiliary (Generator) Room. The common maintenance tasks were developed for administration in a pierside building or van rather than on board a ship.

Finally, a scoring key was prepared in which each behavioral element is dichotomously scored depending upon the presence or absence of the required response.

Observer Training

To ensure that results from hands-on job sample testing are reliable, examiners must be trained to act as unbiased observers of test performance. Therefore, it was necessary to develop a training program for administrators and scorers of the tests.

A training package using a role-playing procedure was developed. The procedure requires that each person to be trained takes a turn playing the role of scorer, examinee, and observer. The role of observer differs from that of scorer in that the former scores the test but does not otherwise interact with the examinee. The rationale underlying this procedure was the expected convergence of scorer behavior as scorers and observers compare and discuss points at which their observations and actions diverge.

Two three-member teams (four contractor and two NPRDC personnel) participated in the training program on board several San Diego based ships. The first training day was devoted to orientation and the second to reviewing standardization procedures, administration instructions and scoring processes. The next six days were spent in role playing and discussion sessions on board a frigate. The training steps were repeated for all job tasks at both work stations as well as at the pierside maintainance test station.

Data Collection

Testing is currently in progress and is being done under contract with a team of four test observers, all of whom are former MM's. Each has approximately twenty years of experience and has attained the minimum paygrade of E7. The remainder of the team, who provide periodic quality control checks, consists of two military members of NPRDC.

Ultimately we will have scores on the hands-on test for about 300 first-term MM's who are serving aboard twenty eight 1052 class frigates. These ships are located at seven test sites including San Diego, Long Beach, Charleston, Norfolk, Pearl Harbor, Newport RI, and Mayport FL.

Performance Test Evaluation

Various analyses will be applied to the data collected. These will include the derivation of distributional characteristics and descriptive statistics, item analyses, and the calculation of various reliability and validity indices. Initially, we are placing a great deal of emphasis on methodology arising from Generalizability Theory.

Performance measures may contain the the same sources of error as traditional paper-and-pencil measures; namely, instability of responses from one occasion to another, non-equivalence of supposedly parallel forms, and heterogeneous subtask responses. Generalizability Theory (Shavelson and Webb, 1981) will be used to estimate the magnitude of each of these error sources, individually and in combinations.

Scores on the hands-on test may result on the one hand from true performance, skill, and ability factors and on the other hand from extraneous features of the testing situation. These features may reflect differences in MM job experience, differences in testing conditions, and differences in MM qualifications across the different testing locations.

By making extraneous variance sources explicit through a generalizability framework, it is possible to identify where error in the criterion measurement is occurring. Consequently, the generalizability framework can act as a conceptual guide to job performance research and increase awareness of potential error sources that need to be controlled or monitored during test administration.

The main questions of interest within the context of a generalizability design involve different possible sources of error variance. The following sources appear most plausible as reliability attenuators:

1. Amount of experience
2. Type of watch station
3. Type of entry level training
4. Location of duty station
5. State of equipment
6. Distractions during testing
7. Test equipment source
8. Scorer differences

These are the facets of the generalizability design for which variance estimates will made.

We are hopeful that we have constructed a sufficiently parallel test that will be objectively observed so that sources of error variance from the administration of the instrument will be negligible. The potential sources of error variance associated with ship and equipment remain to be evaluated.

Summary

The MM hands-on performance test is the first one to be developed within the Navy's job performance research effort. After SME's selected the set of critical tasks to assess technical proficiency of first-term MM's, a hands-on test was developed.

The discovery of informal job specialization during test development led to the construction of parallel test forms for a number of job tasks. A training package for test observers was designed and was used to reduce the variation in test administration and scoring.

Training of the test observers has been completed and data collection has begun on the pre-test ships with two observers at each of three test stations. In order to gather data from 300 MM's on 28 ships in 7 different homeports we will be testing for about one year.

Finally, generalizability models will be used to estimate potential error sources associated with hands-on tests in the overall sample. These sources will include observer disagreement, nonequivalence of parallel forms, unstandardized testing conditions across ships, and the lack of test security. The results of these analyses will be used to guide subsequent analyses.

References

Guion, R. M. (April 1979). Principles of work sample testing: III. Construction and evaluation of work sample tests (ARI TR 79-A10). Alexandria: Army Research Institute.

Kroeker, L. P., & Rafacz, B. A. (November 1983). Classification and Assignment within PRIDE (CLASP): A recruit assignment model (NPRDC Tech. Rep. 84-9). San Diego: Navy Personnel Research and Development Center. (AD-A136 907)

Office of the Assistant Secretary of Defense (M,I, & L). (December 1984). Joint-Service efforts to link Enlistment Standards to Job Performance: Third Annual Report to the House Committee on Appropriations. Washington, D.C.

Shavelson, R. J. and Webb, N. M. Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology 34: 133-166, 1981.

SIMULATION OF HANDS-ON TESTING FOR NAVY MACHINIST'S MATES

Robert Vineberg & John N. Joyner

Human Resources Research Organization
27857 Berwick Drive
Carmel, CA 93923

This paper concerns the substitution of one type of job sample test, a so-called abstracted measure, for another type, a direct measure, in evaluating proficiency. Direct job samples maintain the materials, the responses, and the appearance of actual job tasks. Such samples often encompass entire tasks -- all of the elements of performance that a person might execute together naturally, as in changing a tire or boiling an egg. Even when direct job samples consist only of segments of tasks, each segment generally maintains its integrity and its appearance as one part of a larger unit of performance.

Abstracted job samples, on the other hand, are representations of a process or components of performance that have been disassociated from the elements that naturally accompany them. The appearance of the abstracted component, therefore, is inevitably altered, sometimes very slightly but sometimes radically. An abstracted sample almost always looks more like a test than a job activity, whether or not it is recognizable in terms of its parent task.

An abstracted job sample can be derived from any component or process of task performance that discriminates among performers, regardless of whether the resulting measure has a "hands-on," joblike appearance. An abstracted sample may measure skill or physical ability or job knowledge. Soldering skill required to repair electronic equipment has been evaluated by having persons make solder joints on a "breadboard"; one of the physical requirements for sanitation workers and firemen has been tested by having them lift and carry heavy weights. Perhaps the most familiar and frequently used abstracted job sample is the multiple-choice test of job information.

For evaluating job proficiency, abstracted measures of job knowledge are usually preferred to direct measures because of their efficiency. It is generally accepted that for many tasks a capability to perform can reasonably be inferred if the acquisition and retention of task knowledge have been demonstrated. Yet, direct measures, despite their inefficiency, still have considerable appeal because their use avoids the analysis and alteration of performance inherent in developing an abstracted sample. In the process of translating a task into an abstracted sample, discriminating requirements may be lost and artificial requirements must generally be introduced.

One example is the problem that arises from the use of multiple-choice and other selected-response formats to measure abstracted knowledge: much of actual performance calls for the recall of information, but the multiple-choice format requires only recognition. Ellis, Wulfeck, and their co-workers have designed procedures for checking the adequacy of test formats for reproducing performance requirements in knowledge measurement (Ellis, Wulfeck, & Fredericks, 1979; Ellis & Wulfeck, 1982).

Abstraction of the information-gathering and decision-making components of electronic and mechanical troubleshooting was a popular research area some years ago. The Tab Test (Cornell, Damrin, Saupe, & Crowder, 1954) and like instruments were used as abstract representations of troubleshooting processes. Steinemann (1966) demonstrated, however, that the behaviors elicited in such simulations are quite different from those seen in a direct job sample. Correlations between a hands-on test of troubleshooting and the abstracted measures ranged from about -.5 to +.1. The abstracted measures appear both to invite information gathering that doesn't take place in actual performance (where searching requires much more effort) and to provide considerable structure to a task that is, in actuality, often ambiguous.

Osborn & Ford (1977) reported that pictorial tests provided a poorer measure of knowledge of manual procedures than did multiple-choice tests. They compared performance on four different types of knowledge tests with hands-on performance on tasks like installing a field telephone and disassembling a rifle. In the pictorial test pictures had to be arranged to demonstrate a correct sequence of steps. This seems to have required persons to form a mental image of the entire procedure before arranging the pictures. It appears that the skills required to sort and arrange pictures in the abstracted sample were not part of the actual tasks: average correlations with the hands-on test dropped from the .80s to the high .50s. It is also of interest that persons with lower aptitudes showed their lowest scores when the knowledge tests were based on the picture-arranging format.

I would like to report a different kind of problem encountered in the construction of an abstracted measure for evaluating the proficiency of machinist's mates on Navy frigates (Vineberg, Joyner, & Zimmerman, 1985). In this instance, the abstracted measure was a pictorial simulation of an existing job sample test.

A machinist's mate engages in a variety of sometimes lengthy procedural tasks in the alignment, checking, operating, and troubleshooting of pumps, strainers, condensers, and generators. The number and variety of valves and gauges distributed through the engineering spaces of a frigate present a potentially bewildering array of options to the examinee. In a direct job sample developed earlier, tasks were performed on actual equipment in a ship's engine and generator rooms. In the abstracted job sample, a scenario (identical to that used in the direct job sample) was presented together with photographs of equipment taken in engine and generator rooms.

The examinee was required both to describe actions he would take and to mark the particular equipment components on the photographs that he would observe and manipulate. The photographs had been taken with a wide-angle lens to capture both the equipment relevant to a given task and a good deal of irrelevant equipment in the surround. The pictured equipment provided both a context for recall of task procedures and distractors that could elicit incorrect responses.

In the course of developing the abstracted job measure, it was discovered that the appearance and sometimes the makeup and location of equipment varied among ships of even the same type and class (Knox class frigates) for which the test was developed. Such variability in ship construction is apparently common when ships are built or overhauled at different times, by different manufacturers, in different shipyards. (The arrangement of equipment on a particular ship may sometimes be documented by ad hoc survey, but we are not aware of any compilation of this information across ships.)

The result, of course, is that examinees will vary in familiarity with the pictures of equipment appearing on such a test taken on a particular ship, solely as a consequence of their experiences on other ships with differently configured equipment. The impact of such variation on test performance is likely to be greatest among apprentice job incumbents, for whom the present test was intended, since such persons are least familiar with such variation.

Notice, by contrast, that these variations among ships do not necessarily cause problems in measuring proficiency by direct job sample. In the latter, variations in equipment and work materials in the normal work setting need not lead to differences among examinees in familiarity with a particular set of equipment selected to be represented on the test. Direct job sample tests are often administered in a person's own work situation and use the very same equipment and materials that he or she works with on the job, and standardized score sheets for recording performance can be prepared in a general enough way to be independent of any variations in equipment appearance in the different work settings where the test is administered.

If an abstracted measure, on the other hand, attempts to faithfully render shipboard equipment, as by photograph, the features of the equipment must necessarily be those of particular machinery on a particular ship. Although the consequences of differences in equipment have not been investigated here, the question of test fairness naturally arises when some examinees are presented with familiar materials and others are not.

The effects of such situational variations on abstracted measurement can be avoided by (1) generalizing the abstracted measure so that situational variation is irrelevant or (2) restricting the abstracted measure solely to tasks that use equipment and materials common to all work settings. It may be, of course, that neither of these options is acceptable for making valid inferences about proficiency and that only a direct job sample may remain viable.

REFERENCES

Cornell, F.G., Damrin, D.E., Saupe, J.L., & Crowder, N.A. (1952). Proficiency of Q-24 radar mechanics: III. The Tab Test--a group test of trouble-shooting proficiency (AFPTRC-TR-54-52). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Ellis, J.A., & Wulfeck, W.H. II (1982). Handbook for testing in Navy schools (NPRDC Special Report 83-2). San Diego: Navy Personnel Research and Development Center.

Ellis, J.A., Wulfeck, W.H. II, & Fredericks, P.S. (1979). The Instructional Quality Inventory: II. User's Manual (NPRDC Special Report 79-24). San Diego: Navy Personnel Research and Development Center.

Osborn, W., & Ford, P. (1977). Knowledge tests of manual task procedures. Proceedings of the Annual Military Testing Association Conference, 1977, 634-649.

Steinemann, J.H. (1966). Comparison of performance on analogous simulated and actual troubleshooting tasks (RM SRM 67-1). San Diego: Navy Personnel Research and Development Center.

Vineberg, R., Joyner, J.N., & Zimmerman, R. (1985). Development of a Hands-On Job Sample Test and Scorer-Training Materials for Apprentices in the Machinist's Mate Rating. San Diego: Navy Personnel Research and Development Center.

Inter-Service Transfer of Job Performance Measurement Technology

Capt Jack L. Blackhurst, USAF
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

Herbert George Baker, PhD
Navy Personnel Research and Development Center
San Diego, California 92152-6800

The Department of Defense is coordinating a Joint-Service Job Performance Measurement (JPM) Project in response to a congressional mandate to establish linkages between job performance and enlistment standards. The objectives of this Joint-Service Project are to: (1) develop prototype methodologies for the measurement of job performance; and (2) if feasible, link enlistment standards to on-the-job performance. The overall program will develop measurement techniques for the collection of valid, accurate, and reliable hands-on job performance information that can be related to recruit capabilities. These measures, in turn, will be used as benchmarks against which surrogate indices of performance (less expensive, easier to administer tests and/or existing performance information) will be evaluated as substitutes for the more expensive, labor intensive, hands-on performance measures. The long-term goal of the research and development program is for each Service to establish an operational performance measurement program so that job performance data will be available for use in evaluating personnel and training policies and practices.

Each of the Services is responsible for developing performance measurement technologies on occupational specialties which are comparable across Services. This approach permits the Services to share the technology for similar specialties. A part of the basic strategy of the Joint-Service JPM project is to determine if the technologies developed by one Service can be utilized successfully in other Services. The Air Force is responsible for taking the lead in inter-Service technology transfer. This paper will discuss the transfer to the Navy of a technology developed by the Air Force.

Air Force Testing Technology

The Air Force has developed a comprehensive performance assessment system for the Jet Engine Mechanic Specialty using the combination of Walk-Through Performance Testing (WTPT), rating forms, and related questionnaires. WTPT is a task-level job performance measurement system that combines hands-on task performance and interview procedures to provide a high fidelity measure of an individual's technical job competence. The hands-on component resembles a traditional work sample designed to measure performance on a sample of tasks that have survived the imposition of essential measurement constraints such as testing time/cost or risk of personal injury/equipment damage. The interview component has been added as a means of assessing those tasks that would have been eliminated because of these constraints. Interview testing takes place in the work setting and requires the evaluator to assess an incumbent's proficiency on a task by asking questions designed to uncover knowledge and procedural strengths and weaknesses

related to the performance of that task. The incumbent can answer the questions by a combination of verbal responses, gestures, and demonstration. The interview testing component will be evaluated both as a more cost-effective surrogate and supplement to hands-on measures.

In addition to the WTPT, the Air Force has developed a wide range of rating forms as potential surrogate job performance measures. These include peer, supervisor, and self ratings at four different levels of measurement specificity: task, dimension, global, and Air Force-wide. In addition, a set of questionnaires were also developed to assess job experience and level of motivation. A detailed discussion of the development of the Air Force performance assessment system can be found in Gould and Hedge (1983) and Hedge (1984).

Results to Date

The Air Force developed performance measures for the jet engine mechanic (AFS 426X2) on the three most representative engine types (J-79, J-57, TF-33) currently used by the Air Force. The Air Force, Navy, and Marines all use the J-79 engine in F-4 (fighter) aircraft and have first-term jet engine mechanics who maintain this engine. The Air Force and Navy began discussions in the Summer of 1984 regarding the feasibility of transferring the performance measures developed for the Air Force J-79 jet engine mechanic to the Navy J-79 mechanic. These discussions led to a transfer plan outlining individual Service responsiblities and a three-phase effort to transfer the technology. Phase I was a feasibility study to determine if the transfer could successfully be accomplished. Phase II was the modification of the instruments and procedures, including a pilot test. Phase III is the actual data collection and analysis. The Air Force has responsibility for Phases I and II, with the Navy responsible for Phase III.

To begin the effort, representatives from both Service research laboratories, contractor personnel, Navy training personnel, and two Marine J-79 jet engine mechanic experts attended a workshop at the Air Force Human Resources Laboratory in April 1985. The workshop laid the ground work for the completion of the task. The Marine subject matter experts (SMEs) reviewed the Air Force instruments and test procedures to determine the extent of change needed. Their review indicated that the measures could be transferred with minimal modifications. Following this workshop, several field visits with SMEs at Navy sites were made to confirm the necessary changes, as well as to examine the feasibility of testing Navy mechanics. In addition, appropriate Navy documents (Occupational Survey Report, Training Outlines, etc.) were reviewed to ensure that the tasks represented in the Air Force tests were appropriate for testing Navy and Marine personnel.

A second workshop was held in July 1985 at the Navy Personnel Research and Development Center to receive the contractor's report on the feasibility study, and to determine if the effort should continue. The report indicated that the transfer of performance measurement technology from the Air Force to the Navy and Marine Corps was feasible. The study found that most of the tasks in the Air Force instruments could be used with only minor modifications and that only two of the tasks were not performed by Navy or Marine

personnel due to an equipment difference. For example, the Air Force J-79 engine requires the installation of a starter; however, the Navy or Marine J-79 engine uses an air start and does not have a starter. Therefore, two tasks related to the starter were deleted from the performance test. The rating forms and experience and motivation questionnaires required only minor Service terminology changes and the test logistics did not appear to be a problem. A decision was made to continue the technology transfer effort.

However, a major finding of the feasibility study was that the Navy is phasing out the J-79 engine, which affects the number of first term Navy mechanics that will be available for testing. Because of the limited number of Navy incumbents available, the Navy suggested the inclusion of Marine J-79 jet engine mechanics since, in addition to undergoing identical skill training, they often work side by side with Navy jet engine mechanics on the same engines. Thus, approval has been requested to collect data on Marines, transforming this study into a Tri-Service JPM technology transfer effort.

Another recommendation from the feasibility study was to incorporate the development of Navy job knowledge tests into the study design. The Navy's job knowledge test is different than the typical knowledge test in that it makes extensive use of photographs that the incumbent can use to answer questions. The photographs enable the incumbent to reference the equipment, forms, etc. that he/she would normally use on the job. Inclusion of this additional measurement technique would allow direct comparison of surrogate performance measures developed by different Services on the same sample of incumbents, something which is currently not available in the Joint-Service project. The recommendation was adopted and the measures will be developed for administration with the other performance tests.

Progress has been made in the study. The instruments have been developd or modified as needed and are ready to be pilot tested. Following the pilot test, the measures will be administered to approximately 100 Navy and Marine jet engine mechanics, using test administration procedures similar to those used by the Air Force to collect data on their jet engine mechanics. Results of the data analyses will be available by the Fall of 1986.

Research Benefits

This technology transfer effort will have four major research benefits: (1) This study will serve as a prototype for future attempts to transfer performance measurement technologies. Experience gained from this effort will be invaluable to any future inter-Service performance measurement technology transfers. (2) This is the first attempt at direct comparisons of surrogate techniques, thus underscoring the joint-Service nature of this project. Such comparison of useful Service surrogates provides an expanded assessment of more cost-effective performance measures. The primary surrogate measures from two Services will be directly comparable on the same sample of job incumbents. (3) It allows the Services to gather performance information on additional specialties at significant cost savings because much of the design work has been completed. As a result, the transfer can occur relatively quickly and at a much lower cost than to do the specialty

separately by Service. (4) This effort enhances the total Joint-Service JPM effort. It allows the Services opportunity to share new techniques while enriching their own individual research programs. Tremendous opportunity exists for generation of research ideas and for additional analyses of alternative performance measurement methods. Significant contributions will be made to performance measurement data bases and research.

Summary and Future Research Possibilities

As part of the Joint-Service JPM project, the Air Force and Navy are transferring performance measurement technology developed by the Air Force for jet engine mechanics to the Navy and Marine Corps. The feasibility study and test construction phases have been completed. Data collection will be conducted during the current year. The effort is the first of its kind and will enhance future inter-Service transfer of performance technology. It will provide additional insight into the comparison of performance measurement methods. One very viable research option, a follow-on to this effort, would be for the Services to select a common specialty (e.g., security police or personnel) and develop multiple surrogate measurement techniques for the same sample of incumbents in one Service. To avoid overloading the test incumbent with performance tests, a sample of the various techniques could be used for comparison purposes. This study would allow for direct comparison of all the Service measurement techniques and provide a tremendous data base from which to explore additional research in performance measurement. Such research might include a cost-effective/utility analysis of the various surrogate measurement techniques to determine which technique gives the maximum payoff for the least cost or how surrogates can be combined to provide the greatest amount of performance information.

In summary, benefits derived from this study will have a significant impact on performance measurement research for the individual Services, for the Joint-Service project, and for the field of industrial psychology.

References

Gould, R. B. & Hedge, J. W. (1983). Air Force Job Performance Criterion Development. Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA.

Hedge, J. W. (1984). The Methodology of Walk-Through Performance Testing. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada.

SIMULATION OF INSTRUCTOR AND GROUP PROCESS ROLES WITH MICROCOMPUTER TECHNOLOGY

Barbara L. McCombs
Denver Research Institute

Many students entering military technical training not only are deficient in basic reading skills, study skills, and cognitive strategies, but they also are deficient in motivational skills (McCombs & Dobrovolny, 1982). These motivational deficiencies are reflected in trainees' inability to positively adjust to technical training requirements and implement necessary self-management, personal responsibility, and positive self-control strategies related to self-motivation (McCombs, 1984). Specific deficiencies related to unsatisfactory technical training performance include inadequate goal setting and problem solving skills, self-evaluation and planning skills, strategies for dealing with anxiety and stress, and communication skills (McCombs & Dobrovolny, 1982). A program for remedying these deficiencies, entitled the Motivational Skills Training Program, was developed by McCombs and Dobrovolny (1982) and evaluated with Air Force trainees. The program includes seven self-instructional, printed modules that have been implemented in an instructor-led, small-group format which provides trainees with the opportunity to practice new strategies and skills, share experiences, and develop feelings of rapport with their instructors and peers. Evaluation data indicated that trainees liked the program and found it helpful in their course work and personal lives. Trainees participating in the program also had significantly higher test scores and lower test failure rates than control group trainees (McCombs & Dobrovolny, 1982). Although these evaluation findings with the motivational program pointed to its success, several questions remained. One set of questions concerned training format and whether program cost effectiveness could be enhanced by reducing instructor and/or group interaction requirements through the use of computer-assisted instruction (CAI) for selected portions of the training. These questions were addressed in the reported research, undertaken for the Army Research Institute.

Background. In discussing the use of computer-based media versus conventional media such as instructors or group instruction, Clark (1983, 1984) makes the cogent point that it is not the media per se that influence learning. Rather, it is the content and method of instruction that are critical and the medium is merely an alternative delivery vehicle. Clark argues that in using the medium of computers, one must focus on available instructional theory in finding the necessary instructional methods for fostering the desired learning outcome. He also argues that decisions to use computers are more a matter of implementation issues such as cost, practicality, resources, and equity of access and that this medium can be maximized to address particular implementation problems by focusing on the computer's special features (Clark, 1984). Therefore, in using CAI to simulate critical instructor or group process functions, special delivery features must be carefully matched to content and function requirements.

Some features of CAI that are potentially useful to students who have responded poorly to traditional methods include individualization, mastery learning, and self-pacing. Recent research with these features, particularly the mastery learning model, however, has raised some question about their effectiveness for the disadvantaged background, low ability student (Covington & Omelich, 1981; Federico, 1981; Stinard & Dolphin, 1981; Thompson, 1980). Covington and Omelich (1981) question whether instructional features which include repeated test trials and grading against absolute standards perpetuate a negative failure cycle for low ability/low self-confidence students. On the practical side, however, Siegel and Simutis (1979) point out that within the Army there are many problems associated with providing basic skills training to large numbers of individuals at many different locations (e.g., inconsistent content quality, inconvenient training times,

inappropriate matches of skill levels and basic skills curriculum). Problems such as these led ARI to explore the use of CAI for basic skills training in the Army. In initial studies to evaluate CAI for various types of skills training, Siegel and Simutis (1979) report that CAI was at least as effective as traditional instruction, particularly if instructors are given training in the roles required by the new technology.

An issue of concern in the successful implementation of CAI with a skills training curriculum, then, is the role instructors play in the learning process. At a minimum, it has been suggested that instructors have input into how CAI is used, be given short inservice workshops wherein CAI applications are explained and demonstrated, and receive meaningful role training (e.g., Bloom, 1984; McCombs & Dobrovolny, 1980, 1982; McCombs, Dobrovolny, & Lockhart, 1983; Stasz, Winkler, Shavelson, Robyn, & Feibel, 1984; Swing & Peterson, 1981). In addition, Jernstedt (1983) has argued that individualized computer technologies not be used as replacements for teachers and for group learning, and stresses the need for successful combinations of interpersonal relations and computer technologies. Interpersonal or human functions seen as important include a focus on peer relationships and cooperative goals, and defining leadership roles for students and instructors such that high task engagement results. Computer functions cited as important include frequent and varied active student interaction and the use of visual and other sensory feedback to maintain student attention.

Lubin (1984) argues that for computer-based instruction to be effective, a learning environment has to be created that mirrors the teaching/learning characteristics of live instruction. Similarly, Podenski (1984) argues that this technology should simulate ideal student-teacher interactions and free teachers for more complex tasks, such as diagnosing learning problems, helping students develop appropriate learning strategies, and monitoring instructional effects. Recent advancements now make it possible to include a rich array of audio and visual capabilities within interactive CAI lessons. Ginther (1983) discusses advances in the area of audio/speech devices that can be connected to a variety of common microcomputers. Benefits of these devices include the reduction of reading requirements, the provision of multisensory exploration of new information, and the personalization of materials. Implications for how CAI might be used to simulate group interactions can be drawn from the work of Bloom (1984), Bouton and Garth (1983), Cubberly, Omizo, & Longano (1984), Michaelson (1983), and Neale (1983). These include the use of multiple context case histories of meaningful peer problems in which students can interactively engage in identifying the problem, consequences, and alternative solutions through computer-guided inquiry, imagery, and explanations.

In summary, this selective review has identified features of CAI that can be used to simulate instructor and group process functions. In particular, careful selection of training content, careful design of CAI strategies, the incorporation of personalization through an integration of audio and visual capabilities, the identification of meaningful roles for instructors, and use of the inherently motivating qualities of this medium should contribute to the effectiveness of the CAI enhancements.

Method

Design and development of CAI/audio segments. CAI introductory and practice segments were designed and developed for each of the seven motivational skills modules. A simple computer-controlled audio capability was developed to achieve the personalization desired in the simulation of instructor and group functions. The character "PC," created to simulate instructor functions, was designed to enact three primary roles: facilitator, modeler, and motivator. In the facilitator role, PC helped students acquire new concepts, skills, and strategies via introductory explanations and practice exercises. In the modeler role, PC demonstrated the application of new concepts, skills, and

strategies in guided practice segments. In the motivator role, PC coached and encouraged students to apply new concepts, skills, and strategies in both introductory and practice segments. To provide the identified group functions of peer identification, opportunities for shared problem solving, and peer modeling and feedback, a set of military characters was defined. Case studies and audio scripts were prepared for each character and for the introductory and practice CAI segments. The characters were designed to "grow" as a result of their skill training from an initial inability to solve particular problems to competent problem solvers and self-managers. This transition occurred between PC's guided CAI introductions and CAI practice sessions for each module.

Much of the modeling provided in the CAI segments is accomplished via the audio enhancements, while most of the actual skill practice is provided through the CAI exercises. These two functions are integrated into a single unified presentation via the audio interface in the following ways: (a) the CAI screen reinforces audio information, (b) the audio information precedes a CAI segment or frame; (c) the audio information follows a CAI segment or frame; or (d) the audio is an integral part of a CAI segment or frame. Once the audio tapes had been recorded, pulses were added at points that coincided with CAI screen changes. The contractor- developed audio interface consists of a specially designed interface card which plugs into the Apple IIe game I/O port. The interface receives the pulses from a standard slide-sync audio cassete player. These pulses trigger screen changes and, in turn, allow the CAI software (in this case, the Apple SuperPILOT Authoring System) to control the on/off function of the audio player. This capability allows for computer control of a linear sequence of audio messages that coincide with particular CAI frame sequences, as well as provides for the personalization of skill training introductions and practices, at about one-eighth the cost of videodisc technology.

Experimental design. Six experimental conditions were defined: an historical control group (HC), current control group (CC), a CAI introduction and practice group (CAI), a CAI introduction and instructor practice group (CAII), an instructor introduction and CAI practice group (ICAI), and an instructor introduction and practice group (II). Other independent variables included student scores on the General, Electrical, and Clerical subscales of the Armed Services Vocational Aptitude Battery (ASVAB); military rank; sex; initial judgments of self-efficacy; and initial indices of anxiety and ability to cope with stress. Dependent variables included time to complete the first and second course segments; test failure rates in the first and second course segments; progress index for the entire course; and whether students attrited or graduated. (See McCombs et al, in press, for a description of measures used.)

Subjects. Participants in the study were male and female students in the Electronic Communications (EC) school at Ft. Sill. Students were assigned to one of the five current experimental conditions by designated Ft. Sill personnel, using guidelines and procedures specified by the contractor. At the conclusion of the study, data on a total of 479 students were available for analysis. The number of students in each condition were as follows: 53 in the HC condition, 253 in the CC condition, 55 in the CAI conditon, 61 in the CAII condition, 53 in the ICAI condition, and 57 in the II condition.

Procedures. A 16-hour training program for three Ft. Sill EC course instructors acquainted them with the purpose of the evaluation and the training program, provided guidelines for introducing each module and conducting the small group practice sessions, and described procedures for implementing each experimental condition. Instructors were also trained in the content of the CAI materials and operation of the CAI equipment. A workshop format was used to ensure comparability between the CAI and instructor conditions, instructors were provided with scripts of the case studies applying to each CAI character which they could use as part of their introductions and practice sessions for each module. The pretest measures were administered to students on the day they began

the EC course. Following completion of the pretests, students were assigned to control or experimental groups per contractor procedures. The number of students assigned to experimental conditions in any given class varied from 10 to 15, depending on the class size. The length of training averaged 25 hours and was conducted for 6 hours a day for 4 to 5 days, depending on the rate at which students completed the training. Five Apple IIc systems were available for the CAI conditions. Following the completion of the motivational training, experimental students began their regular EC course work. The EC course is divided into two portions, the first of which is approximately 4 weeks long and the second of which is approximately 5 weeks long, for an average total length of approximately 9 weeks. The course is implemented in a self-paced mode wherein times to complete vary as a function of student ability and motivation. Postmeasures were administered by Ft. Sill personnel in the EC course at the end of the first course portion.

Results and Discussion

Findings. Study results indicated that students who received the motivational skills training with instructor introductions/group practice (II condition) performed better during the subsequent EC course than either students who received no motivational skills training (CC condition) or students who received the motivational skills training via CAI with no instructor practice (CAI and ICAI conditions). This better performance was manifested in significantly fewer test failures and less training time. In addition, as compared to the CC condition, students receiving either the CAI or II conditions tended to have lower attrition rates. Although these differences were not statistically significant, the 8.2 percent reduction for the CAI condition and the 4.4 percent reduction for the II condition may have some practical significance in terms of training costs. For similar students going through the EC course a year earlier (the HC condition) attrition rates were 5.2 percent higher than for students in the current study and progress indices were approximately 12 percent higher. These findings for the HC as compared with current groups were at least in part a function of new procedures implemented in the student battery (housing area) that required mandatory study periods for students not progressing through the EC course at a satisfactory rate. Thus, it is likely that findings with the Motivational Skills Training Program were attenuated in the present study.

Since study findings indicated no overall superiority of the CAI vs. II conditions, some exploratory analyses of potential individual dfferences in subsequent student performance as a function of treatment condition were conducted. These analyses generally indicated that the CAI condition was at least equally effective for approximately half of the students (i.e., those students of high general ASVAB ability and those students with low perception of competence or self-esteem). These findings imply that there may be systematic and reliable individual differences that could be used in differential treatment format assignments, thereby reducing some of the instructor support requirements for the motivational program as well as capitalizing on the use of microcomputer technology for this type of skill training.

Implications of study results. A major assumption of this study was that well designed CAI introductory and practice segments, implemented via a rich microcomputer/audio technology mix, could provide the degree of personalization and simulation of critical instructor and group functions to offset or partially reduce instructor and group process requirements for this type of motivational skills training. Main effect findings with various combinations of CAI and instructor/group supported conditions did not bear out this assumption. This raises the possibility that, in line with Clark's (1983, 1984) arguments concerning the influence of media on learning, that the CAI/audio combination was not sufficiently matched to the content and method of instruction required in this training context. The main effect findings also reinforce

Jernstedt's (1983) point that CAI technologies cannot be used as replacements for teachers and group learning, and that there is a need for a synergetic combination of the human and computer functions to achieve maximum instructional effectiveness.

The exploratory individual difference analyses have suggested, however, that the CAI enhanced version of the motivational training was at least equally effective for some types of students. Of the individual difference variables available for inclusion in these exploratory analyses, the findings with the general ability measure (ASVAB General) are not surprising. That is, a number of studies have found CAI or other multimedia treatments to be as effective as traditional instructor/group methods for high ability students (e.g., Clark, 1984; Kulik, Bangert, & Williams, 1983). On the other hand, the findings that students low in perceptions of competence subsequently perform better if they received the CAI enhanced version vs. the instructor/group version are somewhat puzzling. A plausible explanation, however, may be derived from Bandura's (1982) theory of self-efficacy which suggests that individuals low in perceived competence do not judge themselves as capable of handling particular situations, including interpersonal situations. Because of their low feelings of personal adequacy, they are often threatened in interpersonal situations and fearful of having their perceived inadequacies exposed. For these individuals, then, it is reasonable that the nonhuman medium of CAI may provide a less threatening learning environment, particularly for this type of self-development training that requires considerable self-analysis and self-exposure. As Bowman (1982) has argued, CAI advantages include (a) freedom from fear of reprisal, ridicule, or rejection; and (b) provision for active involvement in tasks that are based on a high probability of success. Thus, it may be that the medium of CAI is a more optimal treatment for those students whose initial perceptions of self-efficacy are low.

The preceding speculations need to be verified by further research. As noted by Siegel and Simutis (1979), CAI's potential lies in its ability to provide individualized, standardized, and efficient instruction, particularly to adult learners who require remedial training. Individualization issues with microcomputer technology thus need to be systematically explored, such that differential assignment to this medium can improve training performance and lead to more cost-effective use of human personnel. There is little question based on the research reported here that instructors and the group process play a critical role in the success of motivational training and the success of individualized computer-based approaches in general (McCombs, in press). Continued research with microcomputer/audio technologies that can simulate instructor and group requirements promises to contribute to a realization of the full benefits of this technology.

References

Bandura, A. (1982). Self-efficacy mechanism in human agency. American Psychologist, 37(2), 122-147.

Bloom, B. S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. Educational Leadership, 41, 4-17.

Bouton, C., & Garth, R. Y. (1983). Students in learning groups: Active learning through conversation. In C. Bouton and R. Y. Garth (Eds.), Learning in groups. San Francisco: Josey-Bass Inc., Publishers.

Bowman, R. F., Jr. (1982). A "Pac-Man" theory of motivation: Tactical implications for classroom instruction. Educational Technology, 22(9), 14-16.

Clark, R. E. (1983). Reconsidering research on learning from media. Review of Educational Research, 53(4), 445-459.

Clark, R. E. (1984, April). Learning from computers: Theoretical problems. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Covington, M. V., & Omelich, C. L. (1981). As failures mount: Affective and cognitive consequences of ability demotion in the classroom. Journal of Educational Psychology, 73(6), 796-808

Cubberly, W. E., Omizo, M. M., & Longano, D. M. (1984, January). The effects of group counseling on self-concept and locus of control among learning disabled children. Paper presented at the annual meeting of the Southwestern Educational Research Association, Dallas.

Federico, P. A. (1981, April). Individual differences and mastery learning in computer-managed instruction. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Ginther, D. W. (1983). Micro's are talking back in the classroom: The promise of speech technology in education. Technological Horizons in Education Journal, 11(2), 105-107.

Jernstedt, G. C. (1983). Computer enhanced collaborative learning: A new technology for education. Technological Horizons in Education Journal, 10(7), 96-101.

Kulik, J., Bangert, R., & Williams, G. (1983). Effects of computer-based teaching on secondary school students. Journal of Educational Psychology, 75(1), 19-26.

Lubin, D. A. (1984). What's right with CBT? Data Training, 3(7), 18-19, 22.

McCombs, B. L. (in press). Instructor and group process roles in computer-based training. Educational Communication and Technology Journal.

McCombs, B. L. (1984). Processes and skills underlying continuing intrinsic motivation to learn: Toward a definition of motivational skills training interventions. Educational Psychologist, 19(4), 199-218.

McCombs, B. L., Bruce, K. L., & Lockhart, K. A. (in press). Enhancements to motivational skill training for military technical training students: Phase I evaluation study report. Alexandria, VA: Army Research Institute.

McCombs, B. L., & Dobrovolny, J. L. (1982, December). Student motivational skill training package: Evaluation for Air Force technical training. (AFHRL-TP-82-31). Lowry AFB, CO: Air Force Human Resources Laboratory.

McCombs, B. L., Dobrovolny, J. L., & Lockhart, K. A. (1983, June). Evaluation of the CMI instructor role training program in the Navy and Air Force. (NPRDC-SR-83-43). San Diego, CA: Navy Personnel Research and Development Center.

Michaelson, L. K. (1983). Team learning in large classes. In C. Bouton and R. Y. Garth (Eds.), Learning in groups. San Francisco: Josey-Bass Inc., Publishers.

Neale, D. C. (1983, April). Specifications for small group activities in instructional designs. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Podenski, R. S. (1984). Implications of electronic learning technology: The future is now! Technological Horizons in Education Journal, 11(8), 118-121.

Siegel, M. A., & Simutis, Z. M. (1979, February). CAI for adult basic skills training. Two applications. Computer based education: Mission of the future, Volume 3, Proceedings of the Annual Convention of the Association for the Development of Computer-Based Instructional Systems, San Diego, CA.

Stasz, C., Winkler, J. D., Shavelson, R. J., Robyn, A. E., & Feibel, W. (1984, April). Staff development for instructional uses of microcomputers: The teacher's perspective. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Stinard, T. A., & Dolphin, W. D. (1981). Which students benefit from self-paced mastery instruction and why. Journal of Educational Psychology, 73(5), 754-763.

Swing, S. R., & Peterson, P. L. (1982). The relationship of student ability and small group interaction to student achievement. American Educational Research Journal, 19(2), 259-274.

Thompson, S. B. (1980). Do individualized mastery and traditional instructional systems yield different course effects in college calculus? American Educational Research Journal, 17(3), 361-375.

ASSESSING M1 TANK COMMANDERS WITH A COMPUTERIZED HAND-HELD TUTOR

Brent Bridgeman
Educational Testing Services
and
Theodore Post
Essex Corporation

Tank commanders on the M1 Abrams tank must be trained to very quickly evaluate a battlefield situation, identify the target which should be engaged first, choose the appropriate weapon and/or ammunition (from among three machine guns and two types of main gun ammunition), issue the appropriate fire command to the loader and gunner, direct the driver to move or change direction if necessary, and maintain communication with other tanks in the unit. For such complex tasks, realistic hands-on training is clearly essential to reach a level of effectiveness, or even survival, on the modern battlefield. But hands-on training is extremely expensive because of high costs of equipment and ammunition as well as the costs of transportation to the few areas that can support full-scale field exercises. In order to make better use of the very limited time for hands-on training, the commander trainees should have already mastered the basic prerequisite skills before going into the field. Thus, for example, if the trainee in the class practices basic elements of a fire command until they are virtually automatic, he is more likely to be able to use them correctly under the multiple pressures in the field environment.

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) has been developing a number of approaches to making traditioinal training more effective. Several projects employ microcomputers combined with videodisks. Such systems are reasonably realistic and are relatively inexpensive compared with hands-on training, but they are still costly enough that the amount of time a given soldier can spend working with the system is limited. In addition, such systems are not easily portable so that soldiers must come to a fixed location at a fixed time to work with the system. Therefore, as a supplement to microcomputer/videodisk technologies, ARI sponsored the development of a low-cost, hand-held computerized Tutor. The intention was to make a device that was low enough in cost (under $150) and small enough in size (no larger than a notebook) that it could be used by soldiers in much the same way that they could use a textbook or manual. The Tutor that was developed as a result of this initiative is a 10" x 11" x 2" device with a 32 character dot matrix display screen on the top, a keyboard on the bottom (numbers 0-9, letters A-E and three operational keys: SAY, ERASE and GO) and an indentation in the center that holds an open 5" x 5" booklet. The use of a printed booklet for the display of test questions, instructional text, and graphics permitted substantial cost savings compared to systems that store this type of textual and graphical information in the computer's memory and display it on a CRT. The Tutor also contains a digitized speech system. (For a more complete description of the Tutor, see Fertner and Bridgeman, 1985).

The Tutor was originally intended to teach technical vocabulary. Its three independent (but mutually supporting) courseware components are — (a) an instructional sequence including a pretest and explanatory text with embedded questions, (b) a drill and practice session (called Word War) in which items answered incorrectly initially are presented again after just one other item has been presented, again after three more items have been presented, etc., (c) a game (called Picture Battle) requiring recognition of an appropriate picture (or portion of a picture) given a spoken stimulus. Because of the success of this approach for vocabulary instruction (Bridgeman and Wisher, 1985), ARI focused attention on other areas where this technology could be applied. One of these areas was instruction in issuing fire commands for M1 tank commanders. A contract was awarded to Educational Testing Service (and its subcontractors Advanced Technology Laboratories and BioTechnology Inc.) to adapt a set of existing instructional booklets for presentation on the Tutor. The remainder of this paper describes how the Tutor's three operational modes (pretest and explanation, Word War, and Picture Battle) were used for fire commands instruction.

Pretest and Explanations

Each of the 28 instructional units begins with a brief multiple-choice pretest. The questions and answer choices are printed in the book, and the soldier responds by pushing one of the A-E keys on the keyboard. After the pretest, the Tutor instantly evaluates the soldiers performance , and allows soldiers with no errors to proceed immediately to the next unit. For soldiers who made errors, the Tutor displays the answer choice they selected followed by the correct response. Thus, this test review becomes the first step in the instructional process.

Next, soldiers who made errors on the pretest are directed to begin reading the explanatory text. Frequent questions are sprinkled throughout the text to ensure that attention and comprehension are maintained. The Tutor provides immediate corrective feedback on these items, but errors should be rare if the soldier is reading carefully.

In the vocabulary module, the "SAY" key was used to make the Tutor pronounce target words that were underlined in the text with code numbers under them. The soldier pushed "SAY" and then the number under the word that he wanted to hear. In the fire commands instruction the use of the "SAY" key is quite different; it is used to provide a brief explanation of incorrect answers. Instead of entering responses on the A-E keys, the soldier makes a selection by pushing "SAY" then 1, 2, 3, or 4. This activates the Tutor's digitized voice system. For example, a page in the book shows a picture of a battlefield scene with several potential targets labeled 1 to 4, and the text asks the student to identify the most dangerous threat. For one incorrect answer. The Tutor says "No, out of range, try again" and for another incorrect answer it says "No, can't kill you, try again." Through this type of oral feedback the soldier learns not only which answers are incorrect, but why they are incorrect. Although this information could be presented through

the display screen instead of the voice system, it would be considerably more distracting, as the soldier would have to repeatedly shift attention from the picture of the battlefield scene to the display screen. Thus, while the use of voice technology in this application was not as crucial as it was for word pronunciation in vocabulary instruction, it still adds a different, valuable dimension to the instructional process.

Word War

The logic behind the increasing ratio review used on Word War (Siegel and DiBello, 1980) applies to many different rote memorization tasks, not just vocabulary learning. In the fire commands module it is used to provide practice in weapon/ammunition selection, and in the identification of the name of threat weapons. For the weapon/ammunition selection routine, the screen first presents a situation (e.g., T-72 at 1000 meters). Followed by three answer choices, presented one at a time (e.g., SABOT, HEAT, M-240). The soldier is instructed to push "GO" when the correct answer appears on the screen. The Word Wars for threat weapon identification were added when data from a preliminary field trial indicated that some soldiers had difficulty reading the text because they did not know, for example, that an SPG-9 is a recoilless anti-tank gun.

Picture Battle

In this game-like activity the Tutor's voice system asks a question based on a picture in the book, and the soldier responds on the keyboard. The display screen is used to keep score. With each correct answer a "projectile" formed by the dots on the display screen moves one step from left to right across the screen. For each incorrect answer an "enemy" projectile moves across the screen in the opposite direction. The object of the game is to destroy the enemy target before the enemy destroys you. Hitting the enemy target is accompanied by sound effects of a shell exploding. For vocabulary instruction, Picture Battle was used to reinforce an association between the spoken name of an object and a pictorial representation of the object (e.g. the Tutor would say "equilibrator" and the solider would find the picture of an equilibrator), eliminating entirely the need to read anything. Although fire commands instruction did not require a task with no reading requirement, the game-like score keeping features of Picture Battle could still be used to good advantage. In one implementation, a battlefield scene is pictured and a situation described in the text; when the Tutor's voice asks "Initial Fire Command" the soldier must select the appropriate fire command for the pictured situation. Immediate corrective feedback is provided and the appropriate projectile advances across the screen, then the soldier is asked to turn to the next battlefield scene and the process is repeated.

Conclusions

As an aid to assessment and training of tank commanders, the Tutor falls on a continuum between traditional paper-based materials and microcomputer/videodisk systems. The Tutor lacks the flexibility and moving graphics capabilities of interactive videodisks, but it is a fraction of the cost of such systems and is easily portable. Although the Tutor is more costly than purely paper-based materials, it provides a much more interesting and interactive environment for assessment and instruction: soldiers may be immediately branched to more difficult material based on pretest scores, visual and auditory explanatory feedback is provided, drill and practice exercises in which the item presentation order is varied depending on student performance are available, and interactions with game-like visual and auditory scoring features are included. Every training technology has advantages and disadvantages, and the hand-held computerized Tutor appears to have a place in the instructional arsenal.

REFERENCES

Bridgeman, B. & Wisher, R. Development of a hand-held computerized vocabulary tutor. Machine-Mediated Learning, 1985, 1(3).

Fertner, K. & Bridgeman, B. Increasing the effectiveness of machine mediated tutoring using embedded testing. Proceedings of the 26th Annual Conference of the Miltary Testing Association, 1984, 75-80.

Siegel, M. & DiBello, L. Optimization of computerized drills: An instructional approach. Paper presented at the Annual Meeting of the American Educational Research Association, April 1980.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily reflect the official policy of the Department of Defense, the Army Research Institute for the Behavioral and Social Sciences, or the U.S. Government.

MICROFORM IN TRAINING

Lieutenant Commander K Jones, BSc., Royal Navy, Staff of the Royal Naval School of Educational and Training Technology.

Lieutenant Commander A E Mizen, BPhil(Ed)., Royal Navy, Staff Officer to Commander in Chief Naval Home Command.

BACKGROUND

1. A major percentage of the technical training of officers and ratings conducted by Naval Establishments has traditionally used hard copy Books of Reference (BRs) and other Technical Publications, supplemented by a series of professionally made training aids. However, the increasing complexity of HM ships has led to a large increase in the documentation required to operate, maintain and repair ships equipments and systems such that a modern frigate carries the equivalent of some quarter million A4 pages weighing some 1½ tons. Microform, having already demonstrated considerable advantages in many areas, has been chosen as the most suitable media for more general naval use, and in particular for technical BRs and publications. Consequently, for training, the withdrawal of hard copy BRs and the introduction of microform requires the conversion of microform to readable print with the aid of readers, projectors and printers.

2. The Royal Naval School of Educational and Training Technology (RNSETT), and the Royal Naval Submarine School (RNSMS), were tasked to identify one or more microform projectors that could operate under the same conditions and to the same standards as conventional overhead projectors, and thereby enable the establishment of a teaching strategy for microform in the classroom. (1)

SCOPE OF THE INVESTIGATION

3. This investigation considered:

a. The type of microform projectors available (forward or rear projection).

b. The applicability of the microform projector to the classroom environment.

(1) Investigation into Microform Projectors, March 1985, Lt Cdr P F Tovey B Ed., RN, and Maj R K Morrison M Ed., RAEC.

c. The most appropriate teaching strategy for use with microform.

d. The range of subject matter to be taught using microform.

CLASSROOM TRIALS

4. Trials were conducted, in the classroom environment, on a range of commercially available microform projectors. Additionally some desk-top readers were also trialled.

5. Each of the projectors was evaluated under similar conditions using an instructor to give a lesson which was observed by:

a. A subject matter expert.

b. An Instructional Techniques officer.

c. A Training Design officer.

d. A Quality Control officer.

e. An officer from the RNSETT Instructional Techniques (IT) Group.

MICROFORM PROJECTION

6. Forward Projection. Each of the microform projectors was evaluated to assess its capability to project an image onto a screen (as a conventional OHP). In all cases the image projected was of poor quality and could only be read by students close to the screen. The edges of the image were blurred, definition was poor and key-stoning presented a problem. After a few minutes there was evidence of eye strain and the instructor found it increasingly difficult to maintain student concentration. This was aggravated by the fact that the projectors could only operate in a darkened room. Used in this mode the image projected was unacceptable to both students and instructor.

7. Rear Projection. At the time of the trial only one projector had a rear projection capability. When used with its A2-size screen it could operate under normal classroom conditions with up to 6 students, comfortably, around the screen at any one time. In this mode the image was clear and eye strain no longer appeared to be a problem. However, the instructor could not teach using the strategy of the lesson and a tutorial style had to be adopted.

8. Compatible Desk-Top Readers. It was found that if a rear projection microform projector was to be used in the classroom, students needed desk-top readers for back-up and consolidation. These readers were operated under the same conditions as the projector.

APPLICABILITY TO THE CLASSROOM

9. The investigation confirmed that microform projectors cannot replace the conventional OHP and should not be considered as such. However, in the rear projection mode they are a most useful training aid and, in this mode, rather than replace the OHP they should be used to complement it.

10. The microform projector should be just another training aid available to the instructor. To be used successfully, however, class numbers would have to be small.

TEACHING STRATEGY

11. As already indicated the rear projection microform projector is suitable only for small class teaching and the tutorial style of instruction would appear to be the most appropriate. This may be an entirely new way of teaching for many instructors. Used in this way the microform projector has the added advantage that the instructor can focus the attention of the students onto the material being presented and as a result have more control of the learning situation (which could never be guaranteed using hard-copy publications).

12. Consequently it is clear that teaching from BRs will no longer be possible, it will be teaching with BRs. Used correctly, therefore, microform may well improve the quality of instruction. Instructors will require training in the use of microform in the classroom and the tutorial method. Properly trained in the correct techniques instructors may overcome some initial qualms about using microform.

SUBJECT MATTER RANGE

13. Microform projectors enable a number of students simultaneous access to BR text and diagrams. It should be remembered, however, that these are photographed pages and as viewgraphs are of the very worst type; as such they are unacceptable. Similarly microform projectors should not be used to present material that can already be done so by using the OHP and viewgraph.

14. The use of microform BRs in training will require the instructor to direct the attention of students to relevant pages of BR text and diagrams. In this situation the instructor is teaching students how to use the microform BRs and as such could be considered a skill. The microform projector is suited to this task although the question of indexing pages, cross-referencing and the requirement to view more than one page at a time will require careful consideration. This may be alleviated by the use of two microform projectors. Other aspects of instruction should be conducted in the traditional manner, emphasising the need for microform projectors to be able to operate under the same conditions as other training aids.

FUTURE DEVELOPMENT

15. Microform projectors are being continually improved. There are now available larger screen rear projection models than the one trialled. There has also been an increase in the number and quality of desk-top readers.

SUMMARY

16. In sum the investigation concluded that:

a. Microform projectors cannot operate under the same conditions or to the same standards as the conventional OHP.

b. Of the microform projectors evaluated, only the rear projection model could operate satisfactorily in the classroom environment.

c. To be used successfully microform projectors must be employed as additional training aids and must be able to operate under the same conditions as OHPs and other aids.

d. The microform projector should only be used by the instructor to direct the attention of students to the relevant pages of the BR text and diagrams. It should not be used to present material that is normally presented by the OHP and viewgraph.

e. Used appropriately, microform projectors could improve the quality of instruction currently undertaken using hard-copy publications.

f. Microform should be used to teach small groups of students and is more suited to the tutorial mode of instruction. For many instructors this will be an entirely new way of teaching and they will require training in the use of microform in the classroom. The RNSETT has already designed an instructional module for 'teaching with microform' as part of their Instructional Techniques Course.

This article is the copyright of the United Kingdom Ministry of Defence © 1985.

Selling Job Analysis

Stanley D. Stephenson
School of Business
Southwest Texas State University
San Marcos, Texas 78666

To the job analysis practitioner, the value of job analysis is patently obvious. Moreover, the value of job analysis data is more and more becoming recognized by others; e. g., witness the increase in articles about various personnel functions that are based on having an accurate job and/or position description, a situation that can only occur when detailed position information is available. There has also been a corresponding increase in the number of books published on the topic of job analysis.

With this growing awareness of the value of job analysis data, one would expect there to be a general eagerness to engage in job analysis projects. However, such is not necessarily the case. When it comes down to releasing the monies to conduct a job analysis (and a job analysis can be expensive), many managers fail to grasp the overall value of job analysis and elect to spend their funds on projects that are more immediate, and perhaps better understood.

The primary problem is that those who truly want the job analysis done (e. g., the trainers, the job evaluators, etc) are not the ones who control the funds; i. e., the manager. On the other hand, the outside consultants, who are contacted to submit proposals to perform a job analysis, initially talk primarily with those who want the work done and not with the managers. However, when it appears that funds are not to be forthcoming, those desiring the project will often arrange a meeting between their manager and the job analysis consultants. Although the meeting is advertised as a meeting of three groups (the manager, the job analysts, and those wanting the project), the meeting often turns out to be a direct discussion between the consultants and the manager whom they have only recently met. Moreover, the manager, since he or she has already indicated that funds may not be available for the project, will be somewhat defensive and will have erected a barrier to counter what is anticipated to be a strong sales pitch. It soon becomes apparent that support from the third party in the meeting (i. e., those who want the project) is non-existent; these individuals are hoping that something (which will not harm their careers within the organization) will happen to change their manager's mind. Consequently, the job analysts not only have to overcome a negative predisposition on the part of the manager, they also have to do it alone.

The underlying problem is that managers usually have but a limited grasp of the value of job analyst data. For instance, they may know about its value in designing a training program but be uninformed about its usefulness in designing a performance evaluation system. Such a situation can be easily understood when it is remembered that usually one function within personnel or training is doing the pushing behind the job analysis request. Their views are presented to the manager, and these views become the sole basis for the manager's perceptions of the value of job analysis data. Consequently, the overwhelming objective for any job analyst making a sales pitch is to break these preconceived

but limited notions and to make the manager aware of the tremendous wealth of information that is available from the collection of job analysis data. Once a manager realizes the indispensable and varied aid that job analysis data can be to the organization, the manager may be more willing to release funds for the project.

This paper briefly reviews the variety of ways in which job analysis data can be of use to an organization. These uses will be discussed with regard to both military and non-military organizations. Next, a suggested approach for expanding an organization's awareness of the value of job analysis data will be demonstrated using actual data.

Uses of Job Analysis Data

Job analysis data are useful in the broad areas of selection, training, production, performance evaluation, promotion, compensation, termination, and termination/retirement. These areas will discussed in the following sections. Not initially obvious military applications will be highlighted.

Selection

Job analysis data allows a detailed job description to be prepared. Such a description is essential to the proper hiring of new employees; if the nature of the job is not known, how can a good hiring decision be made? For instance, research has shown that realistic job previews (RJP) are beneficial in terms of initial hiring and retention; job analysis data lends itself to the preparing of RJPs.

Managers in the military do hire/select personnel to come into their units. In fact many military managers spend considerable time reviewing the files of military personnel. Moreover, military managers go through lengthy interview sessions to hire civilian replacements. The military manager's decision would be greatly enhanced if he or she spend as much time coming to grips with the nature of the job being filled as with the nature of the applicant. However, such information is frequently lacking, and therefore managers spend their time on information that is available; e. g., applicants' personnel folders.

Training

Training can have several purposes. First, it can be used to prepare an individual for entry into the work force. Second, it can be used to maintain proficiency. Third, it can be used to prepare an individual for advancement. In all cases, knowing the job that is currently being done and/or the job to which an individual aspires are crucial to designing and implementing a good training program.

Most military managers realize the first two purposes listed above. However, many managers do not give proper emphasis to preparing their personnel for advancement, often because the "next" job is not understood. Related to this discussion of training is the concept of backup capability. A good manager insures that his personnel have the capability to backup key personnel during periods of illness, vacation, etc. Job analysis data can provide a basis for determining which other position in the organization is the most likely candidate for the backup function. Coming to grips with the concepts of both advancement and backup also permits the military manager to better understand job progression, and in the long term the manager is better able

to improve the morale and productivity of his workers.

Production

Job analysis data can permit the manager to improve productivity in two, general ways. First, it can possibly isolate the differences in task performance of high performers versus low performers. Second, it can provide a picture of how the entire work force "fits" together; i. e., identify the gaps or overlaps in accomplishing the assigned mission. A military manager could make much better work design decisions with job analysis data versus having to rely on direct observations alone.

Performance Evaluation, Promotion, and Compensation

These three factors are directly linked. First, performance must be evaluated. Obviously, a proper evaluation can not be accomplished without knowing the details of the job, yet research has frequently shown that supervisors do not know the details of their subordinates work. Second, promotion is a direct end product of merging the performance in a current job and a knowledge of the next higher level job.

The subtleties of the military promotion system are well known. The military manager who understands the job in question, and has a feel for the best next assignment, is in a much better position to aid his or her subordinates in terms of job progression which in the military directly relates to promotion.

Compensation is a direct function of logically tying salary to worth to the organization; i. e., a classification procedure. The basic method for doing this is to conduct a job evaluation. The need for job analysis data as the basis for any job evaluation is well documented, and most civilian managers are aware of this fact. However, although informal job evaluation does occur, compensation issues do not appear to be a relevant issue for the military.

This brief review demonstrates that job analysis data really is an organization-wide management tool and not just a specific personnel tool. Other aspects not discussed are morale, affirmative action, comparable worth, labor negotiations, and legal aspects in general; job analysis data are also of benefit in these areas.

The bottom line is that when job analysts meet with a manager, he/she needs to be shown how job analysis data can be useful in a larger context than originally thought. It is the "big" picture that will convince a manager that a job analysis project should be funded. Once a project is funded, of course, the specific goals of different personnel/ training functions will also be met.

The above argument should not be dismissed as being appropriate just for the non-military audience. The military manager shares many similarities with his civilian counterpart. Both groups have a vested interest in morale, productivity, training, etc. Moreover, with a steady trend toward converting military to civilian positions, military managers are increasing becoming involved with managing a civilian work force. Consequently, many of the points made in this paper apply to military as well as civilian managers.

Increasing Job Analysis Awareness; An Applied Example

Recently, a job analysis was done of clerical workers employed by a small state agency in Texas; small in the sense that there were only 10 clerical positions surveyed. Job titles were: File Clerk, Receptionist, Clerk, Secretary III, Secretary II, Secretary I, and Office Services Supervisor. This small group of clerical positions has become very useful in demonstrating to others how job analysis can be used in a variety of ways.

Prior to meeting with the management of a different organization, a copy of the clerical task inventory is sent to the organization with a request that a mid-level clerical worker complete it. The completed inventory is processed, and the results are compared to the "population" of 10 workers described in the preceding paragraph. Examples of the utility of job analysis data are demonstrated and discussed with management. An illustration of how this procedure actually works is presented below.

In this case, the inventory was completed by a job incumbent working in a position with the title of Administrative Secretary in the target organization. Task responses produced the following results.

Table 1
Task Cluster Percent Time Spent Results

Task Cluster (Duty)	Percent Time Spent
Written Communications	15.12
Verbal Communications	14.40
Distribution	17.28
Financial	3.60
General Filing	3.60
Use of office machines	3.60
Process general paperwork	0.72
Typing	15.12
Management Assistance	1.44
Miscellaneous	3.60
Office Management	11.52
Forms Management	0.72
Social	3.60
Special - Customer Service	0.72
Special - Dictation	0.72
Special - Notary Public	3.60
Supervision	0.72
Total	100.08

These results could assist management in all of the areas discussed earlier. However, I have found that in a selling session managers like to be shown how job analysis data can be immediately used. Consequently, for the purpose of this paper, two immediate uses are highlighted: backup capability and training for advancement.

Backup Capability

Job analysis data can immediately inform a manager whether or not a key worker's responsibilities can be met when he/she is absent. Comparing this worker's tasks performed with the hypo-

thetical population reveals that for the most part this worker's duties can be covered by other workers should he or she be absent. In some task cluster areas, multiple coverage is available; i. e., more that one other position is performing the identical tasks. Such a situation is convenient for it allows a manager to efficiently spread the workload should the worker in question be absent. The major selling point, however, is that a manager would know exactly where to turn to insure that an absent worker's total responsibilities are covered.

A more detailed analysis reveals that the worker in question performs some unique tasks. In this case, specific task coverage can also be analyzed; analysis results are presented in Table 2.

Table 2
Specific Administrative Secretary
Task Coverage by Other Positions

Task	Covered by:
Edit Administrative Material	Sec II, Supervisor
Write/Draft Minutes	Supervisor
Answer Inquiries	Supervisor
Draft Communications	Clerk, Sec III
Process Payment Vouchers	Clerk
Acknowledge Invitations	Sec III, Receptionist
Conduct Tours	NOT COVERED
Transcribe Dictation	Receptionist
Perform Notary Public Duties	Sec III

Obviously no one other worker can cover these unique and/or critical tasks. Simply put, a manager would have to delegate these tasks to different workers in order to have the work accomplished. Interestingly, for some tasks the best coverage is provided by a Clerk while for others coverage is provided by an Office Services Supervisor. More importantly, there is no one in this population who presently conducts tours. This one piece of information should forewarn a manager about a possible lack of coverage should a key worker be absent. Such information would be perceived as being very useful by any manager.

Training for Advancement

Good managers always look out for their employees, even to the point of insuring that they are prepared for advancement. Job analysis data permits managers to immediately determine the skills needed to perform at a higher job level. For instance, in the example being used, the position in question (Administrative Secretary) would be equivalent to a Secretary II or Secretary III in the comparison population. This comparison was based on a similarity of tasks performed analysis. The next higher job in the population would be Office Services Supervisor. Table 3 presents a summary of the major time spent differences between the present position and the next higher position.

Table 3
Percent Time Spent Differences
Office Services Supervisor Versus Administrative Secretary

Task Cluster (Duty)	Percent Time Spent	
	Administrative Secretary	Office Services Supervisor
Written Communications	15.12	1.01
Verbal Communications	14.40	4.28
Distribution	17.28	3.52
Education/Training	0.00	4.03
Financial	3.60	0.75
General Filing	3.60	0.50
Use of office machines	3.60	1.51
Process general paperwork	0.72	1.01
Typing	15.12	6.55
* Management Assistance	1.44	7.78
Miscellaneous	3.60	2.51
* Office Management	11.52	18.86
* Plan/Organize	0.00	8.55
Forms Management	0.72	8.79
Social	3.60	4.47
Special - Customer Service	0.72	0.75
Special - Dictation	0.72	0.00
Special - Notary Public	3.60	0.00
* Supervision	0.70	20.87
Other:		
Review/prepare documents		1.76
Assist in posting payroll		1.26
Retrieve stored information		1.51
Total	100.08	99.56

* Skills not used in the present position but used in the higher level position.

These results highlight the skills that the administrative secretary should be acquiring while still in his/her present position. Then when an opening occurs, the individual would have the necessary training record to compete for the new position. However, unless someone is aware of the differences between the two positions, a proper on-the-job training program can not be created. Job analysis data is that awareness.

These two examples, backup capability and advancement, illustrate how job analysis data can be put to immediate use by management. Similar examples of how the same data could be used in the other areas discussed earlier could be presented. For instance, a very realistic, realistic job preview could be quickly and accurately created if job analysis data were available. But, the point being made is that job analysis data is not just for training, not just for creating job description, not just for conducting job evaluation, etc. Job analysis data is for managers to use in a wide variety of ways. The better you able to deliver this message, the better will be your chances of having the job analysis program accepted. The approach outlined here will help you do just that.

LINKING WORK, TRAINING, AND PROMOTION IN THE COAST GUARD

Karen N. Jones and John A. Burt
U. S. Coast Guard Institute, Oklahoma City, Oklahoma

The Coast Guard has begun correcting a long-standing problem in its enlisted personnel system -- the mismatch between the work performed by Coast Guard personnel and the content of our training courses and promotion examinations. This mismatch was caused by the lack of an accurate description of the work performed by Coast Guard enlisted personnel.

The work performed by enlisted personnel is described by the enlisted qualifications for advancement (quals). The quals are a series of occupational duty statements describing the jobs performed in each rating (job occupation) at each paygrade. The quals are the basis for the Coast Guard's promotion examinations and rating-specific training. As the primary user of the quals, the training community was aware the quals for many ratings were incomplete and out-of-date. However, the training community did not have the responsibility and authority to correct the problem -- the rating manager has the responsibility for ensuring the quals reflect the minimum occupational standards currently required for the rating.

The rating manager (called rating program manager or program manager in the Coast Guard) for each rating is in the headquarters office responsible for managing the operational program or programs that rating supports. For example, the rating manager for Machinery Technician is in the Office of Engineering. In most instances, a general duty officer is assigned the responsibility for managing a rating or group of ratings as a collateral duty. Since these officers are usually generalists, they often do not have prior training and experience in agency-wide personnel management. In addition, a large established personnel research and development support system (e.g., an on-going occupational analysis program) is not available to assist them.

In 1983, the training community, in conjunction with the rating managers, started a project to correct the mismatch between the work performed by Coast Guard enlisted personnel and the content of the quals, promotion examinations, and training courses. The vehicle chosen to correct the problem was improvement of the promotion examinations (called servicewide examinations within the Coast Guard) which are administered to all eligible enlisted personnel as a part of the promotion competition.

The promotion examinations are designed to test the job knowledge required by personnel in each rating. With the wide diversity of jobs and types of duty stations within each rating, it is feasible to test only a sample of the job knowledge required in a rating. Therefore, identification of what should be tested on an examination is very important. Historically, the content of each examination has been selected by a single subject matter specialist who develops the examination based upon his experience and perception of what was critical job knowledge for the personnel in that rating. After the Coast Guard changed the promotion examinations to pass/fail examinations, it became even more important to identify the knowledges for

each rating which should be tested on the promotion examinations and the knowledges which should be omitted. In this project these knowledges were identified by a panel of subject matter specialists who prioritized the quals for testing on each rating's promotion examination at each paygrade.

Prioritization of the quals was designed to enable the Coast Guard to achieve three goals:

> Strengthen the link between the content of the promotion examinations and the work performed by Coast Guard enlisted personnel.
>
> Develop an accurate description of the work performed.
>
> Develop promotion examinations and training courses reflecting the accurate description of the work performed.

Prioritizing the quals would strengthen the link between the content of the promotion examinations and the work performed. Since meaningful quals were required to get meaningful prioritizations, we incorporated review and revision of the quals as a part of the process. This review and revision of the quals would enable the rating managers to begin developing an accurate description of the work performed (i.e., to revise the quals). Existence of the revised quals would then enable the training community to develop courses and examinations reflecting the actual work performed.

PROCEDURE

During 1983 and 1984, panels of subject matter specialists in 25 ratings met for one week each to review, revise, and prioritize their quals. At the start of the planning for each panel, the rating manager provided a representative for the panel meeting and selected three subject matter specialists from the field. In addition to the rating manager's representative and the three subject matter specialists from the field, the following personnel were on each panel: one subject matter specialist from the resident school, the Institute's subject matter specialist (who developed the promotion examinations and nonresident courses), and a facilitator. Within the panel structure, the program manager's representative chaired the meeting and provided policy guidance, the subject matter specialists provided the rating-specific subject matter expertise, and the facilitator provided testing/training expertise.

Since there is wide diversity within each rating, the subject matter specialists were carefully selected to ensure broad coverage of each rating. Also, efforts were made to include both junior and senior personnel so input from personnel actually doing the work today would be available. The products produced by the panels therefore would reflect both the broad range of experience possessed by senior personnel and the current working knowledge of the first-line supervisors.

Each panel evaluated each qual for currency and adequacy as written, recommended revisions to individual quals, drafted or recommended development of additional quals, and evaluated the adequacy of sections (or content areas) of the quals. Then the panel prioritized each qual for testing on each paygrade's promotion examination. In this prioritization, the panel assigned one of the following evaluations to each qual at each applicable paygrade: (1) essential to the rating, (2) necessary to the rating, or (3) enhancing to the rating. Then the panel recommended testing emphasis for each examination, which equated to number of questions, for each section of the quals. Throughout the process, the panel members were encouraged to identify problems and recommend solutions to the program manager and training community.

Two reports of each panel's evaluations and recommendations were prepared -- one for the Institute and one for the rating manager. The report to the Institute was designed to meet the project's first goal:

> Strengthen the link between the content of the promotion examinations and the work performed by Coast Guard enlisted personnel.

The report for the Institute contained the prioritization of the individual quals for testing on each paygrade's promotion examination. Since this report contained testing guidelines which could be misinterpreted by personnel in the field, it was not given wide distribution.

The report to the rating manager was designed to help the rating manager meet the second goal:

> Develop an accurate description of the work performed.

The rating manager's report contained the problems identified by the panels, the proposed solutions, the overall evaluation of each section of the quals, and the recommendations for testing emphasis on each paygrade's promotion examination. The report also contained the panel's evaluation of the currency and adequacy of individual quals, recommended action for each qual (i.e., retain as written, revise, or delete), and recommendations for additional quals.

IMPACT OF PROJECT

The project's first goal has been achieved and the Coast Guard is working toward achieving the others. The panels prioritized the quals and thus strengthened the link between the content of the promotion examinations and the work performed. The rating managers are reviewing the panels' recommendations and many of them have already revised the quals thus providing the Coast Guard with an accurate description of the work performed. As revised quals become available, the training community is revising the training courses and promotion examinations to reflect this accurate description of the work performed by Coast Guard enlisted personnel.

There have been other tangible and intangible benefits as a result of this project. For example, as expected the subject matter specialists who develop the promotion examinations agreed the prioritizations and testing emphasis reflected the work performed. However, a surprise benefit was unsolicited comments from the field that "the tests were testing what they should be testing." These comments indicated the strengthened link between the content of the examinations and the work performed had been established to an extent where it could be noticed by personnel who were not aware of our efforts to improve the promotion examinations.

Another result was several rating managers using the panels as an opportunity to get additional information. For example:

Some rating managers used the panels to refine proposed quals.

One rating manager adapted the panel meeting to include extra subject matter specialists and an extra task -- generation of a task list for an occupational analysis.

Since the rating managers were involved in planning and conducting the panels, most of them were ready to implement the panels' recommendations upon receipt of the formal report. For many ratings, the panels were able to provide the information needed for the rating manager to revise the quals so additional work was not required. In other cases, additional work was required and the rating managers have been coordinating and funding the work. Examples of this work include:

Occupational analyses using task lists generated by subject matter specialists.

Follow-up panel meeting to refine the quals revised by this project's panel and assist the rating manager perform long-range planning for the rating.

Top-down occupational analysis of all functions performed at Captain of the Port and Marine Safety Offices in enforcing laws, regulations, and Coast Guard policy.

The rating managers are reviewing the problems identified by the panels and the panels' recommendations. Some of the problems identified have already been corrected. For example, in one rating the personnel did not have access to all of the manuals needed to prepare for the promotion examinations. Within a few months, the rating manager had compiled, printed, and distributed an informal publication with extracts from the relevant manuals.

Other tangible and intangible benefits are continually appearing. As mentioned previously, the field is beginning to perceive the strengthened link between the content of the promotion examinations and the work performed. The impact of other intangible benefits (e.g., increased coordination between the resident and nonresident school and greater understanding of the testing/training system) are expected to become apparent over time.

RECOMMENDATIONS FOR SIMILAR PROJECTS

As a result of this project, we have several observations and recommendations which can be useful in planning and conducting this type of project within the Coast Guard or in other organizations.

Panel procedures and materials. Keep your instructions, procedures, definitions, etc. simple. Get the information to the panel members as soon as possible to allow sufficient time for them to review the panel materials. (Our not doing this was the major complaint voiced by the panels.) Word all written materials in terms which are familiar to the reader. This may require different documents for different groups of people (e.g., management or subject matter specialists), but it will save time at the panel meeting and improve the caliber of the products produced. Consider whether or not it is necessary to specify everything in the instructions. We allowed some degree of flexibility for our panels and they had little trouble using the procedures and refining the procedures and definitions to meet the needs of their individual ratings.

Subject matter specialists' perceptions. The subject matter specialists stated the project was exceptionally worthwhile and should be repeated. Their primary concern was that management would not listen to them -- that their recommendations would "get lost" in the system. This concern was handled by assuring them a formal report would be sent to the rating manager. After completion of the panel meetings, an area of frustration was the slowness in implementing the recommended revisions to the quals. If you anticipate a similar problem, minimize the frustration at the beginning by explaining how long it should take to implement the recommendations and why (e.g., the quals manual is revised infrequently and actual printing takes months). In addition, be aware of problems or projects which may interact with the panel's work. As might be expected, when we incorporated previous work or consideration of current problems in the rating into the agenda for the panel meeting, the panel members were more motivated than when previous work (e.g., proposed changes to the quals) or current problems were not included.

Limitations of the process. We did find limits to the process in terms of the type of quals the subject matter specialists could develop. Sometimes the subject matter specialists had difficulty generalizing across different situations. One of the most common responses was: "It varies from unit to unit, district to district, etc." This problem was particularly pronounced when the panels tried to develop quals at the E-7 through E-9 levels. At these levels Coast Guard enlisted personnel are moving from positions with technical emphasis to positions where leadership and management skills are emphasized. The problem we found may be limited to organizations similar to the Coast Guard where duties for a group of people are unit- or even billet-specific.

Facilitator. Most of our panel members were selected because of their position (e.g., Institute's subject matter specialist) or rating-specific experience. The exception to this was the facilitator who could be selected based upon other factors. For this project, the facilitator needed

interviewing skills similar to the skills required of a job analyst to: apply interviewing skills in a group setting; obtain, summarize, and record the consensus of the panel on revisions, recommendations, and evaluations; and ensure participation from all panel members. It was also useful for the facilitator to be familiar with the personnel system, the project's purpose, and the relationship between the products the panel produced and the personnel system. In those instances where we knew the facilitators might not have this type of knowledge, we were able to provide the necessary information in a short orientation session.

Followup. We are in the last part of the third year on this project. During the time required for this type of project, you can expect changes in priorities, funding, and personnel. These changes do not have to stop the project or stop followup activities. The Coast Guard has handled these changes by increased coordination between the rating manager and the training community and among the members of the training community. This has been very important since the rate of change in the quals (and thus changes in our examinations and courses) is much higher than it was a few years ago. To make it easier for your organization, we recommend having one project coordinator throughout the project or, if that is not feasible, maintain clear and concise documentation to ease the transition across personnel.

DETERMINING THE BEST SET OF TRAINING FACTORS FOR USE WITH ARMY OCCUPATIONAL SURVEYS

LAWRENCE A. GOLDMAN, Ph.D. USA Soldier Support Center - NCR

Background. The Army Occupational Survey Program (AOSP) has collected, on a routine basis, Training Factor (TF) data from senior supervisors and managers within enlisted Military Occupational Specialities (MOS) since 1981. The primary purpose of collecting TF data, supplementing information obtained from MOS job incumbents, is to assist training course developers in deciding which tasks should be trained - either by the school which has responsibility for providing formal training or by supervised on-the-job training. Based upon a formal written request to the AOSP by the proponent training school for a given MOS, one or more of the TF included within the Instructional Systems Development (ISD) 8 Factor Model may be used for a specific MOS survey.

These factors are as follows:

(1) Percent of members performing
(2) Average percent of time spent by members performing
(3) Task Learning Difficulty (LD)
(4) Consequences of Inadequate Performance (COIP)
(5) Task Delay Tolerance (TDT)
(6) Probability of Deficient Performance (PDP)
(7) Immediacy of Performance (IM)
(8) Relative Frequency (RF)

In addition, a ninth factor, Training Emphasis (TE), although not part of the ISD 8 Factor model, has been generally used in all AOSP TF surveys.

Despite the wide-spread use of these nine TF. there have been few, if any, previous U.S. Army studies indicating the extent to which these TF are related and the extent to which each of these factors could aid in "critical" task selection. Therefore, this study addressed the following areas:

(1) The degree of commonality between these nine factors; and
(2) Identification of those factors which could best isolate "critical" tasks from non-"critical" tasks.

"Critical" tasks are those found in the Soldier's Manual for each MOS and may be defined as those collective or individual tasks which are required for mission performance in combat or for survival on the battlefield. While there are "critical" tasks for each skill level within a MOS (a skill level corresponds to an authorized paygrade or cluster of authorized paygrades), the focus of this effort was on "critical" tasks required to be performed successfully by entry-level personnel since this is generally the level at which most formal training is provided.

Methodology. The results reported in this phase of the study were based on an Army-wide sample survey of job incumbents and supervisors/managers in the following six MOS: 11B (Infantryman); 63H (Tracked Vehicle Repairer); 91C (Practical Nurse); 76Y (Unit Supply Specialist); 95B (Military Police); and 12B (Combat Engineer). These six MOS were selected for this study primarily because they represent widely different types of work performed by U.S. Army enlisted personnel. Data on percentages of members performing each task and average percent time spent by members performing each task (comprising the first two TF) were collected from job incumbents holding these MOS. The average percent of time spent by all members is based on a 7-point Relative Time Spent scale ranging from "1" (Very much below average) to "7" (Very much above average). Using this scale, a job incumbent rates the total amount of time spent performing each task in his/her present job relative to the time he/she spends performing every other task.

The seven other TF, for which ratings were provided by supervisory/-managerial senior Non-Comissioned Officers (NCO's) in each of these MOS, may be defined as follows:

(1) TE - The amount of emphasis that should be given to each task requiring some type of systematic training, e.g., formal training school, Supervised On-the-Job Training. The scale used ranges from "1" (Very low emphasis) to "7" (Very high emphasis). No response to a task by raters (a value of "0") was presumed to indicate "do not train" and was included in the computation of mean values per task.

(2) LD - Learning difficulty reflects the amount of time required to learn to perform the task satisfactorily. The more time required, the higher the learning difficulty. The scale used ranges from "1" (Extremely low learning difficulty) to "7" (Extremely high learning difficulty).

(3) COIP - This factor relates to the need for identifying tasks essential to job performance, when needed, even if they are seldom performed. The consequence of inadequate performance of certain tasks could result in injury to personnel, loss of life, or damage to equipment. The scale used ranges from "1" (Extremely low consequence - negligible effect on people/equipment/mission) to "7" (Extremely high consequence - may result in injury/death, serious damage to equipment, or failure to accomplish critical mission).

(4) TDT - Task Delay Tolerance relates to how much delay can be tolerated between the time the need for task performance becomes evident and the time that actual performance begins. The scale used ranges from "1" (Extremely low - there is virtually no requirement that an individual be able to perform the task immediately) to "7" (Extremely high - an individual must be able to perform the activity immediately whenever it is encountered).

(5) PDP - This factor relates to insuring that training is given in those essential job skills in which job incumbents frequently perform poorly. It is assumed that for any job, some tasks are more difficult to accomplish than others. The highest ratings on this factor would reflect the most poorly performed tasks. The scale used ranges from "1" (Never performed) to "7" (Very frequently deficiently performed).

(6) IM - This factor is associated with the interval of time between completion of training and the first performance of the task on the job. The scale utilized ranges from "1" (Never performed) to "7" (Initially performed less than 3 months after Advanced Individual Training).

(7) RF - This factor relates to the frequency with which tasks are performed by job incumbents. It ranges from "1" (Very seldom) to "7" (Very frequently).

The sample sizes for the entry skill level personnel (job incumbents) in each of these six MOS follow: (1) MOS 11B - 1424 (skill level 1 corresponding to authorized paygrades E-3 and E-4); (2) MOS 63H - 332 (skill level 1); (3) MOS 91C - 161 (skill level 2 corresponding to authorized paygrade E-5); (4) MOS 76Y - 452 (skill level 1); (5) MOS 95B - 1042 (skill level 1); and (6) MOS 12B - 879 (skill level 1). The number of raters by MOS for each of the seven TF ranged as follows: (1) MOS 11B - 56 for COIP to 104 for IM; (2) MOS 63H - 28 for COIP to 33 for TE; and (3) MOS 91C - 31 for TDT to 43 for PDF; (4) MOS 76Y - 29 for COIP to 82 for RF; (5) MOS 95B - 66 for TDT to 95 for IM; and (6) MOS 12B - 19 for COIP to 30 for TE and IM.

The Comprehensive Occupational Data Analysis Programs (CODAP) were used to obtain data files for the six MOS consisting of computed mean values of the nine TF for each task in the MOS questionnaires. The Statistical Package for the Social Sciences (SPSS) was then used for each MOS. To examine the degree of inter- correlation among the nine TF, a Pearson correlation coefficient matrix was generated followed by factor analysis with varimax rotation of the principal factors. To determine which factors could best descriminate "critical" tasks from non-"critical" tasks, step-wise discriminant function analysis was utilized. For this latter analysis, only the seven TF based on ratings provided by senior supervisory/managerial personnel were examined. This was due to the fact that the senior raters, who could have responded to each task in the questionnaire for each of the seven rating TF, theoretically would have attached greater importance to each "critical" task impacting on successful mission performance. On the other hand, job incumbents would have responded, by direction, only to those tasks which they had performed or had been trained to perform in their current duty position. Also, regardless of the outcome of this study, the two TF obtained from job incumbents would continue to be collected for each MOS to be surveyed. Conversely, focusing on the seven TF rated by senior personnel would facilitate reduction of the number of factors which training development personnel need to examine for "critical" task selection.

Findings.

A. TF Reliability. To determine internal consistency of the data, two types of reliability coefficients were computed for each of the seven TF for these six MOS obtained from senior raters: (1) the average inter-rater reliability of a single rater (r11); and (2) the stepped up reliability coefficient reflecting the overall group of raters for a particular TF (rkk). In general, the r11 values were moderate while the rkk values were consistently high across all three MOS. With respect to MOS 11B, the r11 values ranged from .20 for COIP and TDT to .39 for IM; the rkk values ranged from .91 for COIP to .98 for IM.

With respect to MOS 63H, the r11 values varied from .19 for COIP to .23 for TE; the rkk values ranged from .76 for COIP to .91 for TE. The r11 values for MOS 91C ranged from .15 for TE to .36 for LD; the rkk values ranged from .86 for TE to .95 for IM. The r11 values for MOS 76Y varied from .18 for TE to .31 for COIP; the rkk values ranged from .81 for TE to .93 for PDP. The r11 values for MOS 95B varied from .17 for LD to .39 for IM; the rkk values varied from .86 for LD to .96 for IM. Lastly, the r11 values for MOS 12B varied from .16 for TE and COIP to .35 for RF; the rkk values ranged from .71 for COIP to .92 for IM.

B. Inter-correlations. The Pearson correlation coefficients among the nine factors were all statistically significant for these four MOS: 11B, 63H, 95B and 63H. With respect to MOS 91C, significant correlations were observed for all factors except for LD with TDT and for COIP with the following factors: PDF; RF; % Do (Skill Level 2); and Average Percent Time Spent by all members (Skill Level 2). With respect to MOS 76Y, significant correlations were also noted for all factors with the exception of COIP with RF and average percent time spent (Skill Level 1). Table 1 shows the inter-correlation matrices for MOS 11B, 63H and 91C. Similarly, Table 2 shows the matrices for MOS 76Y, 95B and 76Y.

TABLE 1 - INTER-CORRELATION MATRICES FOR MOS 11B, 63H AND 91C

MOS 11B

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	-.45	.81	.78	.86	.91	.77	.86	.76
LD		1.0	-.25	-.35	-.43	-.58	-.39	-.57	-.55
COIP			1.0	.90	.86	.79	.81	.64	.61
TDT				1.0	.93	.84	.92	.65	.68
PDF					1.0	.92	.94	.76	.76
IM						1.0	.84	.91	.87
RF							1.0	.67	.71
%Do								1.0	.94
AVG % TIME									1.0

MOS 63H

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	.68	.77	.84	.90	.76	.87	.62	.59
LD		1.0	.73	.84	.67	.36	.82	.13	.22
COIP			1.0	.75	.80	.57	.81	.36	.38
TDT				1.0	.82	.56	.90	.37	.42
PDF					1.0	.75	.87	.56	.56
IM						1.0	.66	.79	.75
RF							1.0	.44	.49
%Do								1.0	.94
AVG % TIME									1.0

MOS 91C

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	-.27	.27	.82	.73	.78	.76	.83	.76
LD		1.0	.55	.04	-.40	-.61	-.53	-.47	-.42
COIP			1.0	.50	-.04	-.14	-.08	-.04	-.02
TDT				1.0	.68	.61	.60	.68	.65
PDF					1.0	.81	.83	.82	.77
IM						1.0	.81	.90	.83
RF							1.0	.87	.86
%Do								1.0	.96
AVG % TIME									1.0

TABLE 2 - INTER-CORRELATION MATRICES FOR MOS 76Y, 95B AND 12B

MOS 76Y

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	-.50	.25	.71	.85	.75	.77	.82	.73
LD		1.0	-.20	-.76	-.59	-.78	-.48	-.64	-.63
COIP			1.0	.46	.29	.36	.06	.18	.09
TDT				1.0	.79	.87	.61	.72	.68
PDF					1.0	.84	.76	.82	.76
IM						1.0	.57	.85	.79
RF							1.0	.69	.70
%Do								1.0	.96
AVG % TIME									1.0

MOS 95B

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	-.57	.77	.87	.91	.87	.85	.85	.70
LD		1.0	-.43	-.59	-.61	-.71	-.71	-.57	-.58
COIP			1.0	.89	.68	.60	.66	.54	.42
TDT				1.0	.86	.78	.84	.67	.56
PDF					1.0	.90	.92	.79	.69
IM						1.0	.90	.89	.84
RF							1.0	.79	.76
%Do								1.0	.93
AVG % TIME									1.0

MOS 12B

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	-.52	.71	.83	.78	.74	.64	.71	.64
LD		1.0	-.50	-.58	-.65	-.67	-.63	-.53	-.53
COIP			1.0	.64	.68	.61	.58	.62	.56
TDT				1.0	.80	.77	.72	.56	.52
PDF					1.0	.82	.88	.63	.62
IM						1.0	.77	.77	.73
RF							1.0	.55	.58
%Do								1.0	.96
AVG % TIME									1.0

Based on the generally high inter-correlations of all nine factors for each of these six MOS, it was not altogether surprising that through factor analysis (with varimax rotation of the factor matrices) only two factors emerged for each of the six MOS. As shown in Table 3, the principal factor for MOS 11B (accounting for 90 percent of the total variance) reflected significantly high factor loadings for all factors with the exception of LD. Similarly, the principal factor for MOS 63H (representing 80 percent of the total variance) reflected significantly high loadings for all of the seven TF - provided by senior raters; of some interest was the fact that the two TF produced by job incumbents (% Do and Average percent time spent) had insignificant loadings. The primary factor for MOS 91C (accounting for 78 percent of the total variance) had factor loadings above 0.8 for all factors except LD and COIP. Similar results were also obtained for MOS 76Y, 95B and 12B. As noted in Table 4, the principal factors for MOS 76Y and MOS 95B (each accounting for 90 percent of the variance) reflected significantly high loadings for the nine TF with the exception of COIP for MOS 76Y. The principal factor for MOS 12B (accounting for 89 percent of the variance) also revealed substantially high loadings for each of the nine TF. What these findings indicate is that there is apparently just one underlying general TF, rather than nine different TF.

TABLE 3 - VARIMAX ROTATED FACTOR LOADINGS OF NINE TRAINING FACTORS ON PRINCIPAL FACTORS - MOS 11B, 63H, AND 91C

INDIVIDUAL FACTORS	FACTOR LOADINGS ON PRINCIPAL FACTOR		
	MOS 11B	MOS 63H	MOS 91C
TE	.67	.79	.89
LD	-.16	.89	-.43
COIP	.87	.81	.07
TDT	.92	.89	.80
PDF	.85	.80	.87
IM	.66	.45	.91
RF	.84	.91	.90
% Do	.41	.13	.96
AVG % Time	.44	.20	.91
PERCENT OF TOTAL VARIANCE ACCOUNTED FOR	90	80	78

TABLE 4 - VARIMAX ROTATED FACTOR LOADINGS OF NINE TRAINING FACTORS ON PRINCIPAL FACTORS - MOS 76Y, 95B, AND 12B

INDIVIDUAL FACTORS	FACTOR LOADINGS ON PRINCIPAL FACTOR		
	MOS 76Y	MOS 95B	MOS 12B
TE	.81	.62	.72
LD	-.54	-.55	-.60
COIP	.03	.24	.61
TDT	.58	.38	.84
PDF	.81	.65	.89
IM	.71	.83	.73
RF	.79	.70	.81
% Do	.91	.87	.37
AVG % Time	.91	.91	.36
PERCENT OF TOTAL VARIANCE ACCOUNTED FOR	90	90	89

C. Prediction of "Critical" Tasks. In predicting "critical" tasks from the total task inventory, only the seven TF derived from senior raters for each of these MOS were used as independent (predictor) variables for the reasons noted previously. The objective of the use of stepwise discriminant analysis was to determine the fewest number of TF which best predicted "correct" group membership. That is, it was desired to isolate those predictors which significantly increased the percent of tasks classfied correctly. The first factor entered was that which had the largest value representing the greatest power of discrimination as measured by Rao's V. Subsequent variables selected were those which, when added to the previously selected predictors, measurably increased the percentage of tasks classified correctly.

Table 5 displays those factors which best achieved this objective for MOS 11B, 63H and 91C. Table 6 shows those factors for MOS 76Y, 95B and 12B. For each factor selected in these six MOS, the concomitant change in Rao's V is displayed together with the cumulative percentage of tasks classified correctly.

TABLE 5 - PREDICTION OF "CRITICAL" TASKS - MOS 11B, 63H, AND 91C

STEP ENTERED	INDIVIDUAL FACTOR SELECTED			CHANGE IN RAO'S V			PERCENT OF TASKS CORRECTLY CLASSIFIED		
	11B	63H	91C	11B	63H	91C	11B	63H	91C
1	TE	LD	TE	38.0	195.1	123.5	63.7	76.6	69.0
2	COIP	TE	LD	37.9	61.8	35.9	68.5	80.1	74.1
3	IM	COIP	*	5.9	26.2	*	70.8	82.2	*
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
TOTAL	-	-	-	-	-	-	72.4	83.9	75.2

* The variable for MOS 91C which was entered at step 3 (RF) caused a statistically insignificant change in Rao's V.

TABLE 6 - PREDICTION OF "CRITICAL" TASKS - MOS 76Y, 95B, AND 12B

STEP ENTERED	INDIVIDUAL FACTOR SELECTED			CHANGE IN RAO'S V			PERCENT OF TASKS CORRECTLY CLASSIFIED		
	76Y	95B	12B	76Y	95B	12B	76Y	95B	12B
1	TE	TE	TE	64.9	221.2	48.7	74.0	77.4	66.6
2	COIP	*	TDT	18.5	*	13.2	76.6	*	72.7
3	IM	**	COIP	19.2	**	8.5	80.0	**	73.6
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
TOTAL	-	-	-	-	-	-	80.0	75.6	75.9

* The variable for MOS 95B which was entered at step 2 (PDP) is not shown because when it was added to TE which was previously selected the percent of tasks correctly classified decreased. ** The variable for MOS 95B which was entered at step 3 (RF) caused a statistically insignificant change in Rao's V.

As indicated in Tables 5 and 6, TE was the predominant factor, being entered at either step 1 for five of the six MOS and entered at step 2 for the other MOS (63H). The second best predictor of correctly classfied tasks was COIP, appearing among the top three best predictors in four of the six MOS. Two other predictors, LD and IM, were noted in two of these MOS.

It was observed that the percentage of tasks classfied correctly for MOS 11B, 63H and 95B was higher for "critical" tasks than for non-"critical" tasks. For MOS 11B, these percentages were 80.8 percent vs. 71.4 percent. For MOS 63H, these percentages were 92.0 percent vs. 79.9 percent. Similarly, for MOS 95B, they were 80.2 percent vs. 75.8 percent. On the other hand, the percentage of tasks classified correctly for MOS 91C and MOS 12B was higher for non-"critical" tasks than for "critical" tasks - 79.1 percent vs. 62.4 percent for MOS 91C and 76.2 percent vs. 71.9 percent for MOS 12B, respectively. For MOS 76Y, these percentages were the same - 80 percent.

Conclusions and Implications.

It was evident that there is a remarkably high degree of correlation between the seven TF rated by senior supervisory/managerial personnel and the two TF relating to job incumbent information (reflecting the percent of members performing at the entry skill level and the average percent time spent by these members). What these findings suggest is that rather than nine separate factors there exists, in reality, only one clearly defined factor. similarly, in terms of predicting "critical" vs. non'"critical" tasks, rather than nine individual factors there is essentially only TE (and probably COIP) which could be thought of as generally consistent significant predictors. These finding could provide a basis for training developers to decide which training factor(s) is (are) most beneficial for "critical" task selection and thus facilitate their efforts to identify accurately, entry level training requirements. In turn, the AOSP could collect the most useful amount of TF data, primarily to aid these training developers, from a minimum number of factors. This would enable the AOSP to significantly improve its administration of occupational surveys.

EXAMINATION OF ENVIRONMENTAL DETERMINANTS OF ARMY PERFORMANCE CRITERIA

Darlene M. Olson
U.S. Army Research Institute[1]

Walter C. Borman
Personnel Decisions Research Institute

Job performance has been conceptualized as a product of individual attributes, abilities, and skills which are measurable at the time an individual first enters the organization, of environmental/organizational variables which impact on the individual after job-entry and of the person's motivation to perform. Previous empirical research has investigated work performance in terms of taxonomies of human abilities, values, and personality characteristics (Dunnette, 1976). However, until recently little research has focused on developing taxonomies of environmental/organizational variables or examining relationships between these factors and work-related outcomes.

The major purpose of this research was to examine relationships among individual, organizational/environmental factors, job characteristic variables, and measures of both maximal (e.g., hands-on and job knowledge tests) and typical (e.g., supervisory and peer ratings of performance) performance criteria for first-term soldiers in the Army. This paper discusses results from administering a 110-item Army Work Environment Questionnaire (AWEQ) to 800 first-term enlisted personnel from five military occupational specialties (MOS).

A major impetus for research on environmental variables was the work of Schneider (1978), who proposed that such situational influences as job/task characteristics, organizational practices (e.g., reward system) and climate variables could either directly influence performance or moderate the relationship between cognitive abilities and performance. During the early 1980's several research projects were initiated to develop empirically validated taxonomies of environmental variables (e.g., Peters & O'Connor, 1980; Olson, Borman, Roberson, & Rose, 1984). Results from the development of situational/environmental taxonomies have suggested that situational variables can be identified, categorized and reliably measured. In a series of laboratory studies conducted by Peters and O'Connor, and their colleagues (for a review see Eulberg, O'Connor, Peters & Watson, 1984), results have demonstrated that situational constraints are significantly related to ineffective task performance, job dissatisfaction, and increased frustration.

Although correlational field studies have supported the relationships between environmental/situational variables and affective reactions to the job (e.g., satisfaction), associations between these factors and ratings of performance effectiveness have been inconsistent. For example,

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

O'Connor et al. (1984), reported a weak, but significant relationship between an overall measure of situational constraints and both performance (r = -.12) and turnover (r = .11) criteria in a national sample of convenience store managers. In another study (O'Connor, Peters, Eulberg & Watson, 1984), no significant correlations were observed between situational constraint factors and either performance or reenlistment intentions in seven Air Force jobs. In contrast, findings from other research (Olson, et al., 1984; Steel & Mento, in press) which used different constraint measures have demonstrated significant relationships (rs ranged from .20 to .36 across criteria) between environmental factors and performance ratings.

In general, the mixed results found for relationships between environmental factors and performance suggest that the magnitude of the correlation coefficients are dependent on the level of inhibitors/facilitators actually present in the work environment. Further, the ways situational variables are conceptualized, the kinds of jobs investigated, and the types of performance criteria examined may impact on the observed relationships.

METHOD

Subjects. The research sample contained 800 first-term enlisted personnel from five Army jobs. There were 172 infantrymen (11B MOS), 169 armor crewmen (19E MOS), 144 radio teletype operators (31C MOS), 155 light wheel vehicle mechanics (63B MOS), and 160 medical care specialists (91A MOS). These MOS were sampled at four continental United States and two European Army installations.

Measures. An assessment battery containing an environmental questionnaire and a comprehensive set of typical (e.g., supervisory ratings) and maximal (e.g., job knowledge test) performance measures was used in this research.

Army Work Environment Questionnaire (AWEQ). The Army Work Environment Questionnaire is a 110-item multiple choice instrument that measures 14 dimensions of the Army work environment. The AWEQ was constructed in a two-stage process (Olson, et al. 1984). Briefly, in Stage I, a taxonomy of first-tour environmental influences on soldier performance was derived through application of a critical incident methodology. A total of 282 critical incidents, generated by Army experts (N = 67) and independently content-analyzed by six psychologists, identified environmental/organizational influences beyond the control of the soldier that had a significant impact on performance, either inhibiting or facilitating that performance. The Army work environment taxonomy contains the following nine "job-oriented" factors: (1) Resources/tools/equipment, (2) Workload/Time Availability, (3) Training, (4) Physical Working Conditions, (5) Job-Relevant Information, (6) Job Relevant Authority, (7) Perceived Job Importance, (8) Work Assignment, and (9) Changes in Job Procedures/ Equipment, as well as, the remaining, five "climate-oriented" dimensions: (10) Reward System, (11) Discipline, (12) Individual Support, (13) Job Support/Guidance and (14) Role Models. In Stage II, items were written to cover the content of the 14 environmental dimensions.

Items on the AWEQ are descriptive in nature and respondents are asked to indicate on a 5-point rating scale (e.g., 1 = Very Seldom or Never to 5 = Very Often or Always) how often each environmental situation described in the items occurs on their present job.

Job Performance Measures. The set of typical and maximal performance criteria used in this study was developed as a component of a broader research program conducted under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This comprehensive nine year research effort was initiated to help the Army access, assign, and retain quality personnel.

The typical performance criteria included supervisory and peer job performance ratings. Separate behaviorally-anchored rating scales (BARS), derived from a critical incident job analysis procedure, were used to measure both the MOS (job)-specific and Army-wide components of soldier performance and effectiveness on a 7-point behavior rating format. For each research participant in the five MOS, an Army-wide and MOS-specific rating was computed by averaging the performance ratings across all individual dimensions for supervisors and peers separately.

The maximal performance criteria included hands-on (work sample) tests and job knowledge measures. The hands-on tests for each MOS consisted of 15 tasks identified for the MOS. The individual performance components of each task were scored by trained raters on a pass-fail basis and an overall hands-on score was computed for each soldier by averaging the proportions passed across the tasks tested. Multiple-choice tests were developed to assess job knowledge relevant to each important task for an MOS. An overall job knowledge test score for each research participant was derived as a percentage of the number of items answered correctly.

Procedures. After the supervisor and peer raters were trained to use the Army-wide and MOS-specific BARS, they evaluated the job performance of soldiers in the research sample. Concurrently with these assessments, first-tour soldiers participating in the research were administered: (a) the Army Work Environment Questionnaire and (b) the appropriate job knowledge and hands-on test. For all respondents, scores on the environmental measure were merged with scores from the maximal and typical performance criteria for analyses.

RESULTS AND DISCUSSION

For the total sample, Table 1 presents the means, standard deviations, and reliability coefficients for the research measures. When mean ratings on the AWEQ scale dimensions are collapsed across MOS and installation, results suggest that a complex set of both facilitating and inhibiting influences describe the Army work environment. For instance, the mean ratings for such AWEQ scales as Training ($\underline{M}$ = -3.02), Work Assignment ($\underline{M}$ = -1.90), Reward System ($\underline{M}$ = -1.75), and Job Support ($\underline{M}$ = -1.42) were described somewhat negatively. In contrast, such environmental variables as Perceived Job Importance ($\underline{M}$ = 1.76), Discipline Practices ($\underline{M}$ = 1.10), Individual Support ($\underline{M}$ = .79), and adequacy of Role Models ($\underline{M}$ = .74) were generally described more positively. Uncorrected reliability estimates displayed in Table 1 show that the job knowledge tests tend to be the most reliable of the maximal performance criteria and the Army-wide BARS (supervisors) have the largest coefficients of the typical performance measures. Generally, the AWEQ scale scores, with coefficients ranging from .57 to .78, have adequate reliabilities for a research instrument.

Table 2 presents the intercorrelation matrix for the AWEQ scales. Intercorrelations among the 14 AWEQ scales show that the climate-oriented dimensions are more highly related than the job-oriented factors. For

Table 1

Means, Standard Deviations, and Reliability Coefficients for Selected Measures Across MOS.

Measures	N	M	SD	r[1]
Army-Wide BARS (Peers)	727	4.52	.71	.78-.86
Army-Wide BARS (Supervisors)	722	4.50	.84	.81-.86
MOS-Specific BARS (Peers)	727	4.60	.66	.76-.86
MOS-Specific BARS (Supervisors)	718	4.62	.77	.78-.87
Hands-on Test	685	71.72	16.11	.35-.56
Job Knowledge Test	745	62.47	10.63	.84-.91
AVEQ Scales (# of items):[2]				
Resources (n=7)	734	-.99	4.96	.75
Workload (n=8)	752	-.67	4.34	.58
Training (n=11)	736	-3.02	5.91	.64
Physical Working Conditions (n=6)	741	.67	3.83	.57
Job Authority (n=6)	760	-.25	3.65	.57
Job Information (n=8)	726	.45	4.60	.67
Job Importance (n=7)	725	1.76	4.65	.67
Work Assignment (n=9)	731	-1.90	6.80	.70
Changes in Job Procedures (n=8)	745	-.89	4.21	.58
Reward System (n=7)	736	-1.75	5.14	.78
Discipline (n=6)	751	1.10	4.07	.65
Individual/Support (n=9)	727	.79	5.46	.73
Job Support (n=8)	734	-1.42	5.12	.72
Role Models (n=10)	731	.74	5.98	.71

Note. 1). For performance ratings, the range of interrater reliabilities across MOS are reported.

For Hands-on and Job Knowledge tests, the range of split-half reliabilities across MOS are reported.

For the Environmental scales, Cronbach's alpha coefficients are used as measures of internal consistency.

2). Mean scale scores were computed such that "0" is a neutral environment. Positive mean values indicate positive descriptions of the environment for that scale. Negative scale means indicate the opposite.

Table 2

Scale Intercorrelations for the AVEQ.

AVEQ Scales*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Resources	-													
2. Workload	.52	-												
3. Training	.29	.26	-											
4. Working Conditions	.55	.48	.23	-										
5. Job Authority	.47	.52	.39	.51	-									
6. Job Information	.52	.50	.42	.50	.60	-								
7. Job Importance	.21	.20	.30	.22	.32	.33	-							
8. Work Assignment	.26	.24	.66	.18	.35	.36	.43	-						
9. Job Procedures	.49	.51	.44	.43	.50	.51	.24	.40	-					
10. Reward System	.38	.40	.40	.37	.58	.56	.29	.33	.45	-				
11. Discipline	.31	.31	.18	.37	.47	.48	.30	.14	.36	.45	-			
12. Individual Support	.31	.32	.35	.36	.56	.60	.34	.27	.41	.62	.54	-		
13. Job Support	.39	.40	.44	.38	.64	.62	.34	.37	.48	.73	.48	.72	-	
14. Role Models	.41	.46	.44	.42	.61	.60	.34	.35	.50	.56	.48	.58	.65	-

Note. All AVEQ scale intercorrelations are significant at $p < .05$.

*Correlations significant at $p < .05$.

1). Scales 1-9 are more job-oriented and scales 10-14 are more climate-oriented.

example, Job Support is strongly associated with the Reward System (r = .73), Individual Support (r = .72), and Role Models (r = .65). The Job-Relevant Authority dimension, conceptualized as a job-related factor, has its strongest associations with the climate scales of Role Models (r = .61) and Job-Support (r = .64). Subsequent test development work on the AWEQ, which has included an item-analysis and a principle component factor analysis with a varimax rotation, has been conducted to identify a subset of the original 110 items that best define the factor structure of the AWEQ. Although findings from these analyses corroborate the redundancy displayed in Table 2 for some of the AWEQ scales and tentatively suggest that a five factor solution with 53 items may permit a more parsimonious explanation of the underlying Army work environment constructs, results based on the revised-AWEQ have not been sufficiently cross-validated. Hence, results presented in Table 3 focus on the relationships between the 14 scale scores from the conceptual taxonomy, and a comprehensive set of both ratings of job performance and more objective performance indices.

Table 3 presents the correlation coefficients between the 14 AWEQ scale scores and the set of performance criteria for the total sample. Several interesting findings emerged. First, the largest correlations were found between environmental variables and typical performance measures, specifically the Army-wide BARS. In terms of the number of significant effects, 46.4% of the correlation coefficients between environmental variables and typical measures, as compared with 28.6% of the correlations for maximal criteria, were statistically significant. This difference cannot be attributed to sampling error, since differences in sample sizes for the correlational values shown in Table 3 were relatively minor.

Second, generally the environmental dimensions of (a) Perceived Job Importance, (b) Discipline practices, (c) Individual Support, and (d) the Reward System tended to be significantly correlated with performance criteria for the total sample. In contrast, the AWEQ scale scores on (a) Resources/Tools/Equipment, (b) Workload/Time Availability, (c) Physical Working Conditions, and (d) Changes in Job Procedures/Equipment were not significantly associated with scores on the performance measures. Although the magnitude of these environment-performance relationships are lower than those previously reported with Army field test data from Project A (see Olson et al., 1984), fairly consistent trends have been observed in the pattern of significant relationships between climated-oriented AWEQ scales and performance ratings.

Third, when relationships between typical performance measures and environmental factors were examined, 60% of the correlations between climated-related dimensions and 38.9% of the correlations with job-oriented factors were significantly related to performance ratings. Further, a similar pattern of significant relationships was found between the environmental variables and maximal performance criteria, specifically 50% of the observed correlation coefficients for climate dimensions and 16.7% of the correlations for job dimensions were significantly associated with scores on maximal performance measures. It was predicted that job-oriented environmental factors should have more significant relationships with the objective, maximal performance measures, than the supervisory and peer ratings of overall soldier effectiveness. However, these findings did not support this contention, because a larger percentage of climate-oriented factors than job-oriented factors were significantly correlated with both types of performance indices.

Table 3

Correlations Between AWEQ Scale Scores and Performance Criteria.

Performance Criteria	Scale Scores on Army Work Environment Questionnaire													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Typical Performance Measures														
Army-wide BARS (Peers)	.05	.02	.07	.08*	.13*	.09*	.23*	.07	.04	.11*	.14*	.18*	.14*	.11*
Army-wide BARS (Supervisors)	.01	-.03	.08*	.06	.15*	.11*	.17*	.12*	.01	.11*	.13*	.14*	.13*	.09*
MOS-specific BARS (Peers)	.01	-.04	.06	.05	.07	.04	.16*	.04	0	.05	.10*	.09*	.07	.05
MOS-specific BARS (Supervisors)	-.01	-.08*	.08*	.03	.06	.06	.13*	.11*	-.01	.03	.06	.05	.03	.03
Maximal Performance Measures														
Hands-on Test	-.02	0	-.06	-.01	-.04	.02	.09*	-.02	-.02	-.08*	.08*	.04	-.07	.04
Job Knowledge Test	-.05	-.05	-.05	.03	0	.03	.13*	-.07	-.08*	-.09*	.13*	.11*	-.03	.01

Note. AWEQ SCALES: 1=Resources, 2=Workload, 3=Training, 4=Physical Working Conditions, 5=Job Relevant Authority, 6=Job Relevant Information, 7=Perceived Job Importance, 8=Work Assignment, 9=Changes in Job Procedures, 10=Reward System, 11=Discipline, 12=Individual Support, 13=Job-Related Support, 14=Role Models.

*Correlations which are significant at $p < .05$.

Fourth, consistent relationships were observed between environmental variables and the typical performance measures, specifically the Army-wide BARS, regardless of whether performance was evaluated by supervisors or peers. This finding indicates the existence of some convergence across types of performance criteria with respect to the influence of environmental factors.

Finally, since sampling error may explain many of the observed differences between MOS, only a few potentially meaningful environment-performance relationships will be discussed for the various Army jobs. In the Infantry (11B MOS) and the Medical Specialist (91A MOS) jobs, soldiers' performance on nearly all the typical measures was significantly correlated with Perceived Job Importance (significant rs ranged from .17 to .30). Significant relationships were observed between Work Assignment and the performance of Armor Crewmen (19E MOS) on the MOS-specific BARS and the Hands-on measure. In the 31C MOS, the performance of radio-teletype operators was significantly correlated with Job Relevant Authority (r = .31 observed with Army-wide BARS for supervisors). Performance of mechanics (63B MOS) on the Army-wide BARS (peers) was significantly related (r = .30) to scores on the Individual Support dimension.

Although these specific MOS findings suggest potentially interesting relationships between environmental variables and performance, these correlations are based on substantially smaller sample sizes than those

reported in Table 3 and cannot, without cross-validation, be assumed to represent stable estimates of the true correlational values.

CONCLUSIONS

This research examined correlations between 14 scale scores on an Army Work Environment Questionnaire and measures of both typical and maximal performance. Prior to this applied research in an Army setting, inconsistent findings were reported in the empirical literature with respect to relationships between organizational/environmental variables and performance.

Results from this applied Army research indicated that significant relationships exist between job-oriented and climate-related environmental variables and both job performance ratings (typical measures) and more maximal, objective criteria-job knowledge and hands-on tests. Further, these findings suggest that: (1) environmental factors have their strongest correlations with more typical performance measures such as Army-wide BARS, (2) climate-oriented environmental variables have a larger number of significant effects on maximal performance criteria than job-related environmental dimensions, and (3) generally such job-oriented environmental variables as Resources/Tools/Equipment, Physical Working Conditions, and Changes in Job Procedures/Equipment are not significantly correlated with the comprehensive set of performance criteria.

Perhaps, the weak but significant correlations observed between environmental dimensions and performance may be related to: (1) a lack of sufficiently constraining or facilitating conditions on the part of the environmental variables themselves or (2) contextual factors such as raters adjusting their performance evaluations to compensate for the negative/positive effects of specific work environments.

REFERENCES

Dunnette, M. D. (Ed.) (1976). Handbook of industrial and organizational psychology. Chicago, IL.: Rand McNally.

Eulberg, J. R., O'Connor, E. J., Peters, L. H., & Watson, T. W. (1984). Performance constraints: A selective review of relevant literature. Psychological Documents.

O'Connor, E. J., Peters, L. H., Eulberg, J. R., & Watson, T. W. (1984). Situational constraints in Air Force work settings: Effects on performance, affective reactions and reenlistment plans. Paper presented at the annual meeting of the Academy of Management, Boston, MA.

O'Connor, E. J., Peters, L. H., Pooyan, A., Weekley, J., Frank, B., & Erenkrantz, B. (1984). Situational constraint effects on performance, affective reactions, and turnover: A field replication and extension. Journal of Applied Psychology, 69(4), 663-672.

Olson, D. M., Borman, W. C., Roberson, L., & Rose, S. R. (1984). Relationship between scales on an Army work environment questionnaire and measures of performance. Paper presented at the 92nd annual meeting of the American Psychological Association, Toronto, Canada.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Schneider, B. (1978). Person-situation selection: A review of some ability-situation interaction research. Personnel Psychology, 31, 381-397.

Steel, R. P., & Mento, A. J. (in press). Opportunity knocks: The impact of situational constraints on relationships between job performance criteria. Organizational Behavior and Human Decision Processes.

Initial Standardization of an Air Force Organizational Assessment Survey Instrument*

Lawrence O. Short, Lt Colonel, USAF
Air Force Human Resources Laboratory (AFHRL)

James K. Lowe, Captain, USAF
1605th Civil Engineering Squadron

Janice M. Hightower, Captain, USAF
Detachment 5, Air Force Operational Test and Evaluation Center

In 1981, the Directorate of Research and Analysis of the Leadership and Management Development Center (LMDC) began work on a revision of the primary data gathering instrument used in the LMDC consulting process, the Organizational Assessment Package (OAP). The purpose of this report is to discuss the results of initial standardization research performed on the second generation OAP, now called the Organizational Assessment Survey (OAS). More specifically, the report deals with deriving the OAS factor structure and testing each obtained factor's internal consistency reliability.

In its present form, the OAP survey consists of a computer-scored response sheet and a 109-item (93 attitudinal and 16 demographic) booklet. Responses use a scale of one to seven, with a value of "1" generally indicating strong disagreement or dissatisfaction with the question or statement, and a "7" indicating strong agreement or satisfaction. Through factor analysis, the 93 attitudinal items are combined into factors as presented in Table 1.

Table 1. OAP Factors

Skill Variety	Desired Repetitive Easy Tasks
Task Identity	Advancement/Recognition
Task Significance	Management-Supervision
Job Feedback	Supervisory Communications Climate
Performance Barriers and Blockages	Organizational Communications Climate
Need for Enrichment (Job Desires)	Work Group Effectiveness
Job Performance Goals	Job Satisfaction
Pride	Job Related Training
Task Characteristics	General Organizational Climate
Work Repetition	

Administration of the survey is the first step in the consultation process. The survey is given to a stratified random sample of the organization to which LMDC has been invited. The results of the survey are an important feature in the assessment of task, supervision, climate, and productivity in an organization. The results are handled in a confidential

*All work was completed while authors were members of the Leadership and Management Development Center, Research and Analysis Directorate.

manner between LMDC and the client. After approximately five to six weeks for analysis, consultants return to the organization to provide feedback of data to commanders and supervisors.

When organizational problems are encountered, a consultant and supervisor develop a management action plan designed to resolve the problem at that level of the organization. Within six to nine months, the consulting team returns to readminister the survey instrument as a means to help assess the impact of the consulting process.

The data from each OAP administration effort are stored in a cumulative data base currently containing over 200,000 records for research purposes. These data are aggregated by work group codes developed for this instrument. The data may be recalled by demographics such as personnel category, age, sex, Air Force Specialty Code, pay grade, time in service, and educational level.

The OAP was developed jointly by LMDC and AFHRL at Brooks AFB, Texas (Hendrix & Halverson, 1979a; 1979b). More recently, additional standardization work has been done with the OAP. Short and Hamilton (1981) provided evidence of the factor-by-factor reliability of the instrument considering both internal consistency and test-retest (stability) aspects. In addition, Short and Wilkerson (1981) provided evidence in support of the group differences aspect of OAP construct validity. Webster (1982) also studied construct validity of the leadership and organizational climate areas of the OAP by using a modified multi-trait, multi-method approach, favorably comparing the OAP to the Survey of Organizations (Taylor & Bowers, 1972). Finally, the stability of the OAP factor structure was studied across selected functional area and demographic groups (Hightower & Short, 1982) and across intervals of time (Hightower & Short, 1983). These studies yielded a slightly different factor structure than that currently in use, but showed the revised structure to be extremely stable across all comparison groups. These studies combined with several years of experience using the instrument in the consulting process pointed out ways the OAP could be revised to enhance the process and the accuracy and precision of organizational diagnoses.

Method

Overview and Subjects

The subjects of the OAP-OAS data gathering came from three operational bases in the Continental United States. Since the surveys were administered by LMDC consultants, the test bases were preselected via the consultant's schedule. The data were gathered by LMDC consultants both as part of the consulting process and as an effort to test the new OAS instrument. The subjects responded to both the OAP and the OAS and were then provided the opportunity to verbally express their thoughts concerning the face validity of the OAS instrument. The responses from each person on the two surveys were linked using special coding.

A total of 1583 personnel took the back to back OAP-OAS surveys. The sample consisted of 85.2% males and 14.8% females, compared to the 1984 Air Force ratio of 89% male to 11% female. Ages ranged from 17 to 63, although

96% were younger than 47. The sample consisted of 13% officers, 79% enlisted, 7% General Schedule (GS) civilians, and 1% Wage Board (WB) civilians. The Air Force officer/enlisted ratio is 17% officer, 83% enlisted. Note that if the civilians were removed, the officer/enlisted ratio of the OAS sample would be 14% officer to 86% enlisted. Fifty-four percent (54%) had made either one or two PCS moves, and 24% had been on at least one unaccompanied PCS tour. Over 56% of the personnel had daytime work schedules. Racially, 76% of the sample were Whites, 13% were Blacks (compared to the Air Force's 15%), and 14% were others. Over 27% had more than 12 years in the Air Force, while 21% had less than 2 years of total service time.

Instrumentation

The version of the OAS used for the present study contained 104 attitudinal items for test purposes and ten demographic items in addition to those in the OAP. The OAS survey items were selected to meet two criteria. First, previous factor analyses had shown that the supervisory-related factors and the climate related factors on the present OAP are not separate. Therefore, one criterion was to test new items to create separate factors while eliminating some of the seemingly ambiguous items which presently load with the composite supervisor-climate factor. Second, consulting experience has shown the need for some additional factors not represented in the original OAP. Included in this group are factors pertaining to stress management and intergroup cooperation.

Procedure

Administration. During each survey administration, the OAP was administered prior to the OAS to prevent contamination of OAP results. After completing both surveys, participants were asked to complete a short questionnaire about the OAS. In addition, 384 respondents were randomly selected and verbally polled to see if they had additional comments about the survey. If so, those comments were recorded on the questionnaire.

Factor Structure. Derivation of the factors was accomplished by use of a principal components analysis with a varimax rotation using pairwise deletion. For factor solutions, the "eigenvalue greater than one" criterion was used. In addition, a Scree-test was used to help determine the optimum number of factors to extract. Following Ware, Snyder, and Wright (1976), the Factored Homogeneous Item Dimensions (FHID) criteria were used to assign items to factors. Under these criteria, it was required that all items in a factor have high loadings (+ .40 or greater) only on that factor and low loadings (+ .39 or less) on all other factors in the matrix. This method was a useful check not only for item homogeneity but also for item discriminant validity. An additional requirement imposed was that there be an absolute difference of at least .10 between the primary loading and the item's highest secondary loading. Numeric scores for negatively worded items were reflexed before the factor mean was calculated so that numerically higher factor responses always indicated more favorable responses.

Internal Consistency Reliability. The method of choice here was Cronbach's alpha. Generally, the most popular of the internal consistency methods, alpha can be obtained from a single survey administration and eliminates the inconsistency of splitting items. Its calculation is based on the number of items in a scale or factor and the mean interitem correlation for that same scale or factor. Usually, therefore, as the average interitem correlation or the number of items increases, so does the value of alpha. These procedures must be balanced, however. For example, there is an upper bound on significant increases in alpha from adding items. In addition, adding items that reduce the interitem correlation will not increase alpha. It should also be noted that alpha is often considered the lower bound of internal consistency reliability. Thus, alpha may generally be considered a conservative estimate of the true reliability of a scale or factor (Carmines & Zeller, 1979). Since it is difficult to attach significance levels to alpha, a more direct standard of comparison was used. For purposes of this study, alpha coefficients were considered acceptable at .50 or above, good at .70 or above, and high if .90 or above (Ware, Davies-Avery, & Stewart, 1978; Carmines & Zeller, 1979; Hendrix & Halverson, 1979a).

Results

Factor Structure

The initial factor analysis included 104 attitudinal items of which 47 were new items, 34 were reworded versions of OAP items, 14 were nearly identically worded to OAP items, and 9 were identical to OAP items. Each item was included to help measure one of the factors. Demographic items were not included in the factor analysis.

The initial factor analysis was actually two factor analysis problems, as the SPSS system utilized limited the number of items per run to 100. The initial analysis extracted 15 factors, 13 of which were expected; the remaining two factors accounted for less than 2.5% of the variance and had no items with loadings of 0.40 or higher. Based upon the results of the initial factor analysis, items which did not satisfy the item-loading criteria mentioned in the Procedure section of the report were eliminated from the OAS. Selecting the final items to be included in the OAS was an iterative process to ensure that each item loaded uniquely to its expected factor and to insure that no item that loaded uniquely onto an expected factor was left off the OAS.

The final factor analysis contained only items which satisfied all the item-loading criteria, insuring item "dimensionality." This reduced the OAS from 104 to 77 attitudinal items. Five items from the OAP which dealt with the Need for Job Enrichment were added to the items to be included in the OAS. The Need for Job Enrichment items were included in the final factor analysis even though they had not been field-tested as part of the OAS instrument. We expect these items to continue to load into the same factor once they are included in the OAS. The OAS survey contained five Combat Readiness items which, even though they loaded strongly into one factor, were removed from the OAS survey to be included instead within LMDC's Combat Attitude Survey. Only 11 of the final 77 attitudinal items had means above 5.5.

The 13 final factors are listed in Table 2. The factors accounted for 68.5% of the variance. We had hoped that the items relating to Recognition would load into a separate factor; however, these items loaded with the Organizational Climate items. Two of the extracted factors, Advancement and Intergroup Cooperation, consisted of three items and the Work Conditions factor consisted of two items. The factor loadings of these items onto their respective factors were all above 0.54.

Table 2. Derived OAS Factors

Job Goals	Supervisor
Job Characteristics	Advancement
Task Autonomy	Organizational Climate
Training	Intergroup Cooperation
Work Support	Work Group Effectiveness
Work Conditions	Need for Job Enrichment
Effective Stress Management	

Internal Consistency Reliability

The 77 attitudinal items which loaded onto each of the 13 factors were used to determine Cronbach's alpha for each factor. Values of Cronbach's alpha ranged from .71 to .96, with 10 of the 13 factors having alphas of at least .84. The factors with three or fewer items, Advancement, Intergroup Cooperation, and Work Conditions had alpha values of .75, .71, and .69, respectively. Even though the three small factors had fewer items than desired, their reliability scores were acceptable.

A Final Comment

The present research, then, shows the OAS factor structure to be statistically sound and generally consistent with the literature as well as previous experience with similar items on the OAP. The factors also showed very high internal consistency reliability. The combination of these two findings provides initial support for the OAS as either a consulting or survey research instrument. The OAS seems appropriate to replace the OAP as it provides a more solid, replicable factor structure with fewer items and factors than the OAP.

References

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Beverly Hills, CA: Sage.

Hendrix, W. H., & Halverson, V. B. (1979a). Organizational survey assessment package for Air Force organizations (AFHRL-TR-78-93). Brooks AFB, TX: Air Force Human Resources Laboratory.

Hendrix, W. H., & Halverson, V. B. (1979b). Situational factor identification in Air Force organizations (AFHRL-TR-79-10). Brooks AFB, TX: Air Force Human Resources Laboratory.

Hightower, J. M., & Short, L. O. (1982). Stability of the Organizational Assessment Package factorial validity across groups. Paper presented at the 90th Annual Convention of the American Psychological Association, Washington, DC.

Hightower, J. M., & Short L. O. (1983). Temporal stability of the Organizational Assessment Package factor structure. Paper presented at the 91st Annual Convention of the American Psychological Association, Anaheim, CA.

Mahr, T. A. (1982). Manual for the Organizational Assessment Package survey (Report Number 82-2560). Maxwell AFB, AL: Air Command and Staff College.

Short, L. O., & Hamilton, K. L. (1981). An examination of the reliability of the Organizational Assessment Package (LMDC-TR-81-2). Maxwell AFB, AL: Leadership and Management Development Center.

Short, L. O., & Wilkerson, D. A. (1981). An examination of the group differences aspect of the construct validity of the Organizational Assessment Package. Proceedings of the 23d Annual Conference of the Military Testing Association. Washington, DC: Military Testing Association, 1981.

Taylor, J. C., & Bowers, D. G. (1972). Survey of organizations. Ann Arbor: Institute for Social Research, University of Michigan.

Ware, J. E., Jr., Snyder, M. K., & Wright, W. R. (1976). Development and validation of scales to measure patient satisfaction with health care services: Volume I of a final report. Part A: Review of the literature, overview of methods, and results regarding construction of scales. Publication No. PB-288-329. Springfield, VA: National Technical Information Service.

Ware, J. E., Jr., Davies-Avery, A., & Stewart, A. L. (1978). The measurement and meaning of patient satisfaction. Health and Medical Care Services Review, 1(1), 1-15.

Webster, A. H. (1982). An assessment of the construct validity of the leadership and Management Development Center's Organizational Assessment Package (Report number 82-2640). Maxwell AFB, AL: Air Command and Staff College.

PRODUCTIVITY MEASUREMENT: ISSUES AND CHALLENGES

Paul van Rijn, Ph.D.

US Army Research Institute[1]
5001 Eisenhower Avenue
Alexandria, Virginia 22333-5600

This paper describes some of the lessons learned during the Army Research Institute's (ARI) efforts to conduct an independent outside evaluation of the outcome of a macro-level sociotechnical system (STS) intervention at a large Army maintenance Depot. This particular intervention focused on a major division of the Depot which consisted of about 900 workers, mostly Army civilians in skilled blue-collar occupations, and which had primary responsibility for the repair and overhaul of the airframe (as opposed to the engine and transmission) of Army helicopters.

The STS analysis and the development of the design recommendations were conducted by two outside STS experts working closely with a team of 12 Depot volunteers who had been carefully selected to represent both the different levels and occupations within the Depot. This phase of the intervention lasted nearly one year and involved: (1) the development of a Depot Philosophy Statement, (2) the explicit articulation of the mission of the Depot, (3) the identification of the key variances (deviations from the norm) affecting the work of the Depot and their controls, (4) an analysis of the social (people-related) system of the Depot, and finally, (5) the development of 12 specific recommendations for change.

Although ARI's preliminary efforts to evaluate the intervention outcome were initiated from the beginning of the STS analysis, these efforts were not a prominent component in the discussions of the STS analysis or in the development of the recommendations. Goals and expectations for increased productivity and enhanced quality of working life were expressed, but seldom in operational or measurable terms. ARI's role was that of independent outside evaluator.

Toward the end of the STS analysis, when the change recommendations had been developed, this researcher first became responsible for ARI's evaluation of the intervention. Numerous archival measures had been identified previously as potential indicators of increased productivity and a survey instrument had

---

[1]The views expressed in this paper are those of the author and do not necessarily reflect the views of the US Army Research Institute or the Department of the Army.

been developed to assess key organizational variables. Implementation of the recommendations had not yet been begun and the conditions for tracking the intervention before, during, and after implementation seemed optimal. However, appearances were deceptive.

First, it became apparent that the archival measures identified were much more complex than had previously been assumed. Interpretations of the simplest measures, such as "Number of aircraft produced," surfaced a long list of questions. Did the count include all aircraft or only some models? Were crash-damaged aircraft with their special requirements included in the counts? And to what extent did the overhaul of non-Army aircraft and special projects affect the measure?

It was also apparent that, with few exceptions, the "number of aircraft produced" matched exactly the number that had been contracted each month. If 30 aircraft were required, 30 were produced. If 28, then 28. Overproduction would result in the aircraft being "carried over" into the next month or would result in diverting work to neglected shopfloor maintenance or other support tasks. If production was behind schedule, the remedy might be to authorize more overtime or schedule extra shifts until the schedule was met. This often resulted in a flurry of activity toward the end of the month.

Other measures posed similar challenges, not the least of which was that for every measure there were multiple reporting systems. The printout of productivity report titles alone was over one-half inch thick. Nearly everything was counted and logged into the highly automated reporting systems. There were no simple ways for sorting through these productivity reports to determine which reports would be most useful for a meaningful evaluation of this intervention. There was no single expert who could advise on the utility of each report, many of which were produced solely to comply with reporting requirements and formats imposed from outside the Depot.

Due to the multiple reporting systems, it was not unusual to find a measure as non-complex as "sick leave useage" produce inconsistent or discrepant results. Often the data would be different because they were computed differently. For example, on one report, "sick leave" was the rate per employee per 100 work hours, while on another report it was the rate per pay period, or 80 work hours. Even the term "monthly" took on different meanings. On different report it might refer to the varying number of calendar days, the available number of workdays, or simply two pay periods without regard to the number of actual workdays involved. Seldom were these variations obvious from the report titles and painstaking investigations were required to resolve even the most obvious discrepancies. These individuals responsible for maintaining the various reports were often not aware of these discrepancies, since they maintained only one type of report. And although they were

highly dedicated, those maintaining the printouts were not always the same people who knew the precise details of the derivation and meaning of the numbers being reported.

A second major measurement issue that emerged was the level of analysis. It has already been suggested that global measures, such as "number of aircraft produced," are often difficult to interpret and may mask real productivity gains made at more molecular levels. Figure 1 shows some of these levels and the types of subtasks that are required to process a helicopter airframe through its 18 overhaul stations. Over 60 relatively independent work centers (not all are shown) are involved in maintenance and overhaul of the helicopter airframe. Each work center has its function, from disassembly to electronics repair, to painting and flight test.

Reducing the specificity of measurement to this level might be intuitively attractive were it not for the tremendous volume of data that would somehow need to be collected, analyzed, and synthesized. This would represent not only be a collosal task for the researcher, but, more importantly, it would place considerable unexpected strain on the resources of the organization to duplicate and make all this data available.

Besides the volume of data that would be generated by a work center level of analysis, the researcher would now also have to face the challenge of comparing and aggragating data across work centers with very diverse technical processes and different outputs. Except for a few measures, such as sick leave, there were few productivity indicators that could be aggragated meaningfully across work centers. In addition, it raises the question of how a researcher can assess the productivity of one work center when the productivity of that work center depends significantly on the productivity of one or more other work centers? The paint shop, for example, cannot be expected to demonstrate high productivity, if it is not provided sufficient numbers of airframes to paint or if those airframes all arrive at the same time.

Other major issues are the reliability and validity of the measures themselves. Even some of the more promising measures proved on closer scrutiny to be highly variable from month to month and were likely to be of dubious quality for research or evaluation purposes. The large monthly variations could often be traced to unevenness in the submissions of data for the report, individual differences in the criteria used to report the data (e.g., What is a reportable defect?), computer system failures or software upgrading, undocumented changes in the way the data are recorded or calculated, re-establishment of the engineering standards or norms from which performance efficiency ratios are calculated, and so on.

The timing of the evaluation measurement is also critical. In an intervention as large and complex as that of the Army Depot, it was difficult to specify at what point in time a measure is

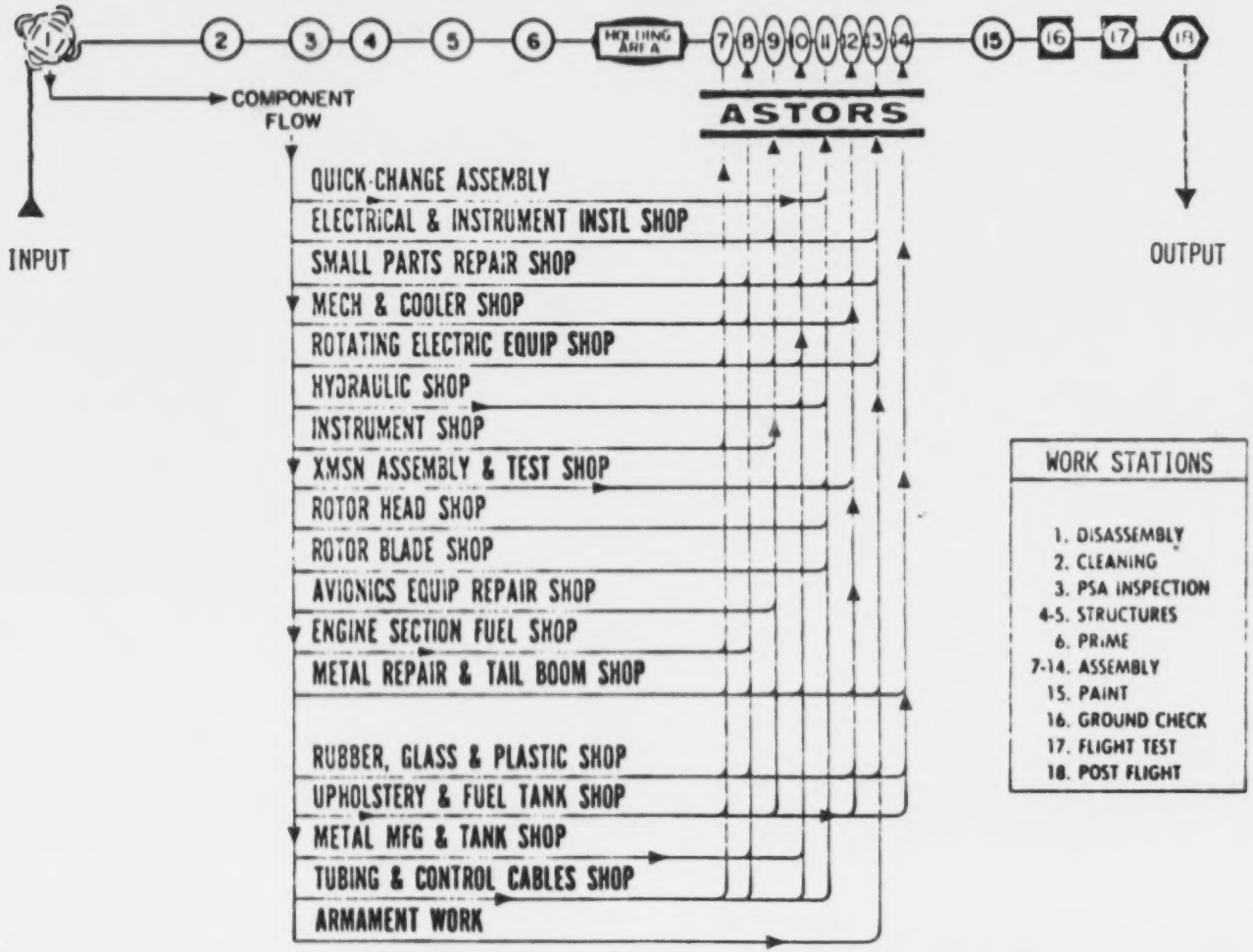


Figure 1. Schematic diagram of the workflow through the 18 work stations of the Airframe Division. "ASTORS" is the Automated Storage and Retrieval System.

before, during, or after implementation. This was due to the fact that different recommendations were implemented at different times and had different durations. Training, for example, can be expected to result in an early decrease in productivity with its benefits not becoming fully realized for years. Other recommendations, however, can have a more specific and immediate impact. Finally, it can be argued that implementation started the moment the analysis phase was begun -- long before the recommendations were implemented or even developed.

Another important measurement issue is the identification and control of confounding variables. It is highly unlikely that during an intervention of any magnitude that there are not also other factors that impact on the measures used in the evaluation. For example, during the intervention at the Army Depot, the Depot went into a massive hiring mode. Does this increase productivity? Or, does it decrease productivity as productive workers have to divert some of their attention to the training of the new and unskilled hires? Merely identifying these potential confounding factors is a difficult task, but this is not nearly as difficult as assessing the direction and magnitude of the effect, or of determining how the effect varies over time and from work center to work center.

Finally, the survey instrument, developed to study Depot organizational characteristics that might contribute to an effective intervention, proved to be insensitive to the changes that were ultimately implemented. This was because the survey focused largely on the structure and work processes of the Depot; while the intervention, as a whole, did not significantly impact on these areas. Given the nature of the organization, major organizational or process changes would have been highly unlikely. Consequently, it should come as no surprise that there were virtually no changes detected over the 14-months timeframe in which the survey was administered. The few questions that did demonstrate an attitude change (some over 20 percentage points) tended to be questions that related directly to the philosophy and principles of the STS approach to organizational change.

Lessons Learned

Based on the experiences with this particular intervention, a number of productivity measurement lessons were learned:

1. The measurement of an organization is likely to be much more complex than it initially appears.

2. Independent outside assessment of an organizational intervention is counterproductive. First, it does not focus the assessor on the most meaningful measures and it does not provide for a meaningful reduction of the multiple measures that are available.

3. Interpretation of the numbers alone, without full knowledge of the process from which they derive and the context in which they occur is likely to be highly misleading.

4. The assessor must work very closely with the organization -- from the very beginning of the intervention -- and assist the organization in identifying the measures that will be used to assess the success of the intervention. This requires defining the goals and expectations of the organization and the outcomes of the change recommendations in terms of measurable outputs.

5. To the extent possible, the assessor and the organization should be co-investigators and partners in the intervention, mutually learning about real organizational concerns and both being concerned that the data are meaningful and trustworthy and that the inferences derived from the data are logically sound.

6. To the extent possible, the different stakeholders in the intervention need to be identified early and their criteria for judging the pluses and minuses resulting from the intervention must be articulated.

7. Searching for unitary cause and effect relationships is inefffective. The assessor and organization must work together to clarify the major interdependencies and contexts in which measurement occurs.

8. The measurement aspects of an intervention need to be an integral part of every phase of the intervention. To the extent that positive outcomes can be identified early, the momentum of the intervention is likely to be sustained. It may be advisable to deliberately include in an intervention some recommendations that are likely to have an early demonstrable payoff. The more these payoffs can be expressed in terms of "hard" dollars saved or productivity gains, the more likely the intervention will continue to receive the top-management support it needs to remain resourced.

9. Finally, resources required to collaboratively assess an intervention are not trivial but are an essential and important investment, not just for the evaluation of the particular intervention, but also for the continued monitoring of an organization's productivity and quality of working life.

A MORALE AND MISSION RELATING STRATEGY

Raymond O. Waldkoetter
U.S. Army Soldier Support Institute
Fort Benjamin Harrison, Indiana 46216-5060

The Mission Area Analysis (MAA) assesses the long-range capability of a programed force to perform required combat tasks. The analysis is designed to discover task deficiencies and correct them with changes or solutions in doctrine, organization, training, and materiel. The MAA process also provides a basis for applying advanced technology to future Army operations with the inherent aim of increasing combat effectiveness. The MAA is performed by a study group at the Training and Doctrine Command (TRADOC, 1985) Centers or schools. There are 13 MAAs, projected to be conducted in groups of three once every four years, and are usually initiated prior to the research, development, and acquisition cycle. Throughout an MAA's analytic process, the soldier must be considered as an integral system. Deficiencies in combat effectiveness must be remedied either by improving soldier performance directly or by changing doctrine, organization, training, and/or materiel. In any case, the soldier is the key element in the combat equation.

The individual soldier must be represented as a "combat system" integrating TRADOC combat development activities, relating to improving individual performance with human technologies or equipment that will enhance individual capabilities during combat (Weisz, 1980). Soldier factor issues are identified during the MAA process to advise or assist proponent TRADOC combat developers in solving problems related to soldier motivation, capability, and performance. Direct assistance and advice are given to proponents so that the focus is on consideration of soldier factors as the means to enhance individual and unit performance. The Soldier Support Center (SSC) works with proponents in developing MAAs and study advisory follow-through to assure soldier factor consideration is also carried over into the materiel acquisition process (MAP).

The purpose of soldier factor consideration is to: insure that potentially critical soldier factors are considered in the MAA process; provide a guide for considering soldier factors in the MAA process; and promote efforts to assess the impact of soldier factors on MAA deficiency solutions. Study efforts in the MAA process should specifically include consideration of soldier factors. After-the-fact consideration of soldier factors is largely inefficient. The individual soldier should not be regarded as an add-on to a materiel system. Soldiers must have a defined role within the mission area. Serious consideration of soldier factors can reveal ways to avoid substantial degradation and enhance ability to perform the mission.

The views and opinions expressed in this paper are those of the author and should not be taken as an official policy of the Department of the Army.

METHOD

Soldier factor issues become critical and must be remedied to the extent that they detract from the soldier's job and combat performance. Soldier factors affecting morale and mission performance must be defined in some scheme to begin to understand how they may affect duty or combat. Historically, commanders have relied on the human/soldier dimension or "moral force" to decide conflicts when their forces were equivalent in other respects (Zais, 1985). Knowing relationships probably exist between the global notion of morale and mission success has not led to any readily reducible process to confirm these relationships. Since factors related to morale and mission success are so numerous, it seems best to devise a logical scheme to analyze "some," appearing to have noticeable affects.

Soldier factors, then, are defined in this paper to be such behavioral determinants as cohesion, stress, values or ethics, mission (sense of), and performance (sustained). Studied in the MAA context, these soldier factors are also seen as being affected by three primary determinants - systems/ weapons, individual/group characteristics, and leader/management actions - adding to the comprehensive soldier factor affects. Morale as a collective state-of-mind can be identified by the interaction of these soldier factors when producing a positive attitude expressing motivation and satisfaction (Motowidlo, Dowell, Hopp, Borman, Johnson & Dunnette, 1976). Thus, the fully adequate consideration of soldier factors also conveys the basis for understanding and assessing the level of unit soldier factors as morale.

Soldier performance is affected to a lesser or greater extent as soldier factors are used as corrective actions. Materiel has given limits, (i.e., tensile strength and payload), and consideration must be given to reliability, availability, and maintenance. Soldiers also have their given limits, (i.e., strength, endurance, and morale); and again, consideration must be given to "responsibility, availability, and maintenance" (fitness and readiness). A thorough consideration of soldier factors will likely guide the planning for and improvement of combat performance beyond the MAA process. The concept of soldier factors - cohesion, stress, values or ethics, sense of mission, and sustained performance - may be developed under the primary determinants with descriptive action-task (AT) statements suggested for each factor to improve mission analysis and corrective action design.

Each proposed corrective action or solution to an identified task deficiency is derived from concepts related to doctrine, training, organization, and/or materiel. A recommended solution following the soldier factor review will succeed only to the degree that it remains compatible with other projected changes and is supported within the given resource constraints. The perfect solution is an elusive goal and the best alternative may demand tradeoffs in deciding which soldier factors should have the most

attention in obtaining a solution. As the MAA proponent task force or assisting soldier factor analyst examines the interface between solutions and soldier factors, the following procedures serve to implement consideration of soldier factors in the MAA process. Consideration of the soldier factors has to identify and determine for proposed solutions: action-task (AT) statements that will affect soldier combat effectiveness; task conditions and standards; the combination of soldier factor considerations that can most likely affect task and combat success; any expected changes in soldier factors related to proposed solutions; and, essential elements of analysis and measures of effectiveness for soldier factors. The probable improvement in critical tasks is determined in a proposed solution by considering the extent to which those negatively contributing factors can be reduced or countered, and positively contributing factors can be augmented and supported for corrective action. The total impact of soldier factors is evaluated and determined through the: impact of the proposed solutions on soldier factors; impact of the soldier factors on proposed solutions; soldier factor initiatives, constraints, and limitations in proposed solutions; and, acceptability and feasibility of implementing the proposed solutions.

RESULTS AND DISCUSSION

The analysis proceeds from evaluating the impact of proposed solutions on the most critical soldier factors to questioning how aspects of proposed solutions will affect one or more of the soldier factors and vice versa. For example, will a proposed solution "require leaders to increase training to cope with increased intensity of battle (stress)?" Action-task (AT) statements like this are used to judge the affordability and supportability of a solution. The AT statements can be used as operational task statements and prioritized in terms of importance for the proponent mission. Priority for mission accomplishment can be indicated by having a subject-matter panel judge whether they disagree or agree with the level of relevance of an action toward solving mission area needs. Then any deficiencies associated with the prioritized action can be identified and remedied by the most appropriate types of solution. A second way the proponent may utilize AT statements is by prioritizing actions most to least pertinent. Thirdly, if a deficiency would occur in any primary factor set, judges or raters can indicate to what degree it would increase the probability of inadequate performance. As corrective actions/solutions are reviewed in relation to selected soldier factor actions or issues, then the proposed solutions for doctrine, training, organization, and/or materiel will be defined and prescribed accordingly. If a solution is constrained by any soldier factors it may have to be altered.

Although numerous AT statements can be generated, in most cases 20 or less will suffice if an in-depth review is conducted. Some 105 AT statements were generated with 21 for each factor and seven for each factor under a given primary determinant. Any variation of AT statements is possible depending on

the mission area concerns. Table 1 is presented as a working example using an analytical scheme based on occupational-task analysis procedures.

TABLE 1

	SOLDIER FACTOR CONSIDERATIONS	
Systems/Weapons (Parameters/capabilities of systems/ weapons, how to employ, and what to expect from adequate use and effects.)	Individual/Group Characteristics (Physical health, skilled task performance, skilled interpersonal performance, integrity/mental health; group goals, skills, endurance, and efficiency.)	Leader/Management Actions (Refined mix of goals, behaviors, power, and style used to insure effective combat operations and productive effort.)
A. Cohesion (Sense of belonging, feeling a part of something, sharing of problems, or bonding together of soldiers and leaders to accomplish any purpose resolved.)		
1. Assure system/weapon training is related to suitable skill level.	1. Estimate whether personnel can function in more than one (1) job.	1. Communicate clear leader goals and objectives for survival.
2. Determine operational indicators to improve combat assignment.	2. Monitor impact of probable personnel actions on troop confidence.	2. Set proper examples in combat exercises and combat.
3. Analyze need to coach or add more skill for system/weapon team.	3. Identify special skills to cope with unusual problems or situations.	3. Initiate motivating rather than coercive L/M actions.
4. Identify expected operational measures to sustain efficiency.	4. Develop attitudes to support troop loyalty.	4. Review and apply cohesive practices and information.
5. Specify most helpful actions to support operator/teams.	5. Support job satisfaction interests in preparing personnel assignments.	5. Take part in a cross-section of activities to help official direction.
6. Project critical time allocation for major system/weapon functions.	6. Define duties to take advantage of available skills.	6. Supervise and participate in expected standards of excellence.
7. Assess strengths and constraints in achieving proficient operations.	7. Recognize examples of esprit and sacrifice in battle.	7. Demonstrate reliability and compassion in crises.
B. Stress (Reaction of the mind/body to extreme demands with sensations of tension and anxiety, and which must be managed to avoid degraded performance.)		
1. Confirm the adequacy of communications for system/weapon selected to be deployed.	1. Analyze combat exposure levels and parity among troops.	1. Define procedures to manage stress continually in action.

Essentially, the impact of proposed solutions on soldier factors is found by indicating which factors are expected to be affected, with some idea of the magnitude of consequences anticipated. It may be adequate to merely identify the soldier factors affected and indicate the order of severity. A more detailed interrogative analysis may be desirable to define just what, how, and why, the factors are posing as unresolved and critical soldier factor issues. The impact of soldier factors on proposed solutions can be approached by identifying those factors which can alter constraints or limitations in solutions. These factors, if augmented or supported, can act as a "force multiplier" for continuous combat operations.

Conducting a productive consideration of soldier factors engages as much detailed analysis as a proponent can afford. An iterative consensus is recommended to relate the magnitude of soldier factor and AT statement judgments to primary determinants and mission success. Numerous questions are required in regard to mission areas to assess the importance of soldier factors and the interaction among such factors. Each proponent will have to decide on the scope and level of resolution needed to deal with soldier factor issues in the given mission area. Corrective actions and solutions for task deficiencies require detailed analysis to audit their development from task deficiency priority to type of solution (i.e., training). A speculative scenario will allow the analyst to relate the probability of mission success to morale by progressively assessing a primary determinant, soldier factors, and AT statements, as these indicate a need for corrective action.

The analyst or subject-matter panel can select one of the primary determinants for guiding soldier factor analysis (systems/weapons, individual/group characteristics, leader/management Actions). A value or rank of importance may be assigned in comparison with the other determinants, e.g., 1, 2, or 3. Next, the analysis process can estimate the magnitude or criticality for each of the five soldier factors toward assuring adequate combat effectiveness under the given determinant and threat scenario. If the five factors would not appear "critical to adequacy" to the analyst or panel, the scale magnitude would tend to be at the lower scale values. That is on a scale of 1 to 5, the relative priority of factors would tend to vary from 1 to 3. Now, the analyst or panel can select the AT statements considered most useful or needed to favorably modify the five soldier factors by augmenting or enhancing performance and related perception. Additional or new AT statements or items can be constructed as may be necessary. More detailed AT items might be prepared under each chosen item to bring about the consideration or desired improvement through one or more corrective actions or soldier factor solutions. Next, the magnitude or criticality would be estimated toward enabling the major AT item or subitem to be performed, i.e., by rating on the probability of inadequate performance on a scale of 1 to 5 (low to high).

The final phase of analysis is to derive a combined magnitude of the three levels - primary determinant, soldier factor, and AT item/subitem - to estimate the criticality expected for some composite probability of inadequate performance (Stark & Waldkoetter, 1985), specifying needed change or solution for mission success. The three-level estimate of criticality and multiplicative/additive combining of each level can yield a conditional/notional index of priority with analyst/panel consensus for each of the primary determinants. As the most simplistic example it could suffice to base a conditional analysis on a single primary determinant rather than all three. If Leader/Management Actions was ranked highest for a value of 3, the highest rated soldier factor with a value of 5 and one AT item alone rated for a value of 5 a product of 75 is obtained. Then, if the other remaining ranked factors

(1 to 4) have only one AT item each rated with values of 5, a sum of 50 would result to add with the 75 for a total of 125. The lowest possible overall total is 15 following such a forced solution. A scale mid-point of 70 and above may be arbitrarily given as the hypothetical area in which serious indications have morale implications and are noted for the probability of inadequate performance. Further, if Leader/Management Actions in Table I was considered of highest combat importance, the Cohesion factor was ranked highest for combat effectiveness (criticality), and only one AT item, 1. Communicate clear leader goals and objectives for survival, was given the highest probability for inadequate performance, then the analysis process would have yielded an indication of trouble in need of some corrective action. That action being to make certain the communication task was mandated in doctrine, training, organization, and/or materiel to build cohesive performance utilizing leader action to insure the expected resolution.

This pre-R&D overview offers a procedural basis for soldier factor (SF) consideration that will logically and systematically identify potential SF constraints and needs related to soldier impact on combat effectiveness. Analytical consensus and review should strongly influence this procedural model as a notional strategy. In this approach it seems logical to infer that if leader, cohesion, and communication actions are expected to be impaired by critically inadequate performance, then morale and mission success are inexorably related. Some principle of proportionality must apply in that any task action for either demands a complementary action for the other to achieve a dynamic balance.

REFERENCES

- Motowidlo, S.J., Dowell, B.E., Hopp, M.A., Borman, W.C., Johnson, P.D. & Dunnette, M.D. (1976). Motivation, satisfaction, and morale in Army careers: A review of theory and measurement. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Stark, R.W. & Waldkoetter, R.O. (1985). Tactical communications Mission Area Analysis: Human fatigue and stress study (draft). Fort Gordon, GA: U.S. Army Signal Center.

- U.S. Army Training and Doctrine Command (TRADOC). (1985). Mission Area Analysis Handbook (draft). Fort Monroe, VA: Headquarters, Deputy Chief of Staff for Combat Developments.

- Weisz, J.D. (1980). Human factors engineering in research, development, and acquisition. Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

- Zais, M.M. (1985). Ardant du Picq: Unsung giant of military theory. Army, 35, 56-64.

COGENT

A MICROCOMPUTER BASED CRITERION REFERENCED TESTING SYSTEM

DALE KERKMAN
TRAINING DEVELOPMENT UNIT
NAVAL TRAINING CENTER
GREAT LAKES, ILLINOIS

COGENT is a microcomputer based academic management tool that offers a workable, reliable, and efficient solution to the implementation and administration of criterion referenced testing. It provides a means of generating printed tests and evaluating student responses to not only each test but also to distinct content areas within each test. These content areas can be related to the objectives within the test. COGENT provides detailed test results which indicate mastery of the course objectives as well as student tracking and record keeping.

COGENT has a built in test authoring system that enables a person to easily enter test items into an item bank. An editing function allows the test author to make changes to an item, delete an item, insert an item, or move an item to another position within the item bank. Each test can be subdivided into up to 10 content areas. Each content area contains all of the test items for one discrete concept (terminal objective, enabling objective, lesson topic, etc). In addition, the content area parameters, which are used in the printing of the test and in determining if criterion was met for each content area, are easily entered into the system with the author program. Up to 100 test items may be entered into each distinct content area.

A test printing function enables the operator to easily print a hard copy test. Pre-defined parameters determine how many items are selected from each of the content area item banks. A maximum of 100 items can be selected from the content area item banks and printed. The test items can be randomly selected by the computer or manually selected by the operator. The test items can be printed in random or sequential order. The choices for each of the test items can also be printed in random or sequential order. If computer selected and random printing of test items and choices is chosen when printing new tests each test printed by COGENT will be unique from any other and will assist in minimizing the possibility of test compromise. Multiple choice and true/false items are accommodated in the test item banks.

COGENT will significantly decrease the number of manhours presently required to prepare, revise, type and print new examinations. New variations of a test can be generated as often as desired.

The tests are scored from optically scanned answer sheets. COGENT analyzes not only the entire test but also individual content areas to determine if each student mastered the subject matter contained in the content areas. Test results are printed after the last answer sheet has been scanned. The report lists all of the students who took the test along with their associated overall test scores. It also includes for each student a score for each of the content areas and a notation indicating any failed content areas. Individualized prescriptive assignments can also be printed for those students requiring remediation and retesting. Customized remedial tests for each of the students requiring retesting can also be printed. These tests will be made up of only those items from the failed content area.

COGENT allows students to be transferred from one class to another with all test data remaining intact. Students can also be put into a 'hold' status and then transferred to an active class. In addition, students can be disenrolled from school. Any student that is transferred, dropped or put into a hold status requires a corresponding code to be entered so that reports can be printed listing any and all students by any desired code.

COGENT provides the following reports:

*Class Roster - provides an alphabetical listing of the students including name, rate, SSN, and other pertinent student data.

*Student Cumulative Summary - provides a listing of all the students along with their overall score on each test. Included is a notation indicating the students' pass/remedial status for each test. This notation indicates the number of times each student was retested in order to achieve criterion on all of the content areas for each test. This report also includes each student's cumulative average and the class cumulative average.

*Content Area Performance Summary - prints a detailed analysis of each student's performance on each test. It includes the overall test score, the score for each content area the first time the test is given and any/all scores on subsequent remedial tests. It lists the class performance statistics such as: number of students taking the test; class test average; number of students passing all content areas on first attempt; number of student requiring remediation after first attempt; etc.

*Student Ranking - provides a printout of the class with the students ranked high to low based on each student's cumulative average. In addition, those students meeting honor graduate requirements can be recognized with an appropriate notation.

*Academic Review Board Worksheet - printed for those students who experience academic difficulties. It is used by the Academic Review Board in determining academic action.

*Student Memo - form printed for students who are assigned remedial study/testing. It lists the time and location where the remedial study/testing will take place and is used to assist in managing the remedial study and retesting program.

*List of Transfers/Drops - provides a printout of the students who were either transferred to another class, or disenrolled from school. It also includes the transfer/drop code and the date the action took place. This report can be printed for one student, all students, any specified transfer/drop code, or all transfer/drops codes.

*Item Analysis - provides a D&V item analysis and also a response count analysis which are used to determine the effectiveness of the test items. It also includes the mean, median, and standard deviation for the test.

Some of the benefits that can be realized by COGENT are:

1. Ability to identify specific problem areas within a test:

COGENT grades the answer sheets and evaluates each test by content area (objectives). This enables the proctor/instructor to identify problem areas and remediate to those specific areas. For example, with conventional testing a test covering subject matter such as Introduction to the Naval Chain of Command; Naval Terminology; AC Theory; DC Theory; and Safety Precautions would be evaluated on a students overall test score. It is very likely that a student could pass the test with an overall score of 70% but incorrectly answer the items on the Safety Precautions subject matter. With COGENT each of the individual subject matter areas (content areas) are evaluated and a report is generated listing all of the students and their scores in all content areas. The areas where the student failed to meet criteria are annotated so that individual remediation can be prescribed.

2. Reduction in remediation/retest time:

Without COGENT, if a student fails a test, he/she is remediated and retested on the subject matter that is covered by the entire test. Since specific problem areas within each test are readily identified by COGENT, the student is remediated and retested ONLY on the specific areas where he/she is experiencing difficulty. COGENT can automatically generate a content area specific remedial test that is tailored to the problem area/s of each student.

3. Reduction of errors in grading tests:

All tests are computer scored so that the possibility of an error in grading is greatly reduced.

4. Reduction in time to update/add/change items in the test item banks:

The authoring program allows the test manager to easily add or delete items in the test item bank. Also, changes can be made to existing items. For test audit purposes, each test item has the corresponding objective annotated along with the data each item was entered in to the item bank.

5. Reduction in time to validate test items:

The item analysis option allows the test manager to easily generate statistics which can be use to determine item difficulty and validity.

6. Reduction in time to print a new test:

The time to print a new test is reduced from hours to minutes because the computer automatically prints the tests from predetermined data.

7. Reduction of compromised tests:

All test items can be randomly drawn from a bank of items. The items can then be printed in random order with the distractors also randomized. This significantly reduces the possibility of test compromise.

8. Reduction of student/class management time:

COGENT maintains a record of all student performance/grades and provides various reports by individual or for an entire class. COGENT also provides a roster and a class ranking report with notation for prospective honor graduates. Printed forms for the Academic Review Board for those experiencing academic difficulties can also be printed. Students can also be transferred from one class to another with all grades remaining intact.

9. Ability to effectively administer a criterion referenced curriculum:

Without the automatic scoring and evaluation of tests, it would be impossible to attempt to administer an effective criterion referenced curriculum. COGENT automates all of the functions necessary to perform these tasks which results in a more highly skilled graduate.

COGENT is currently operating on an Apple computer with a 10 megabyte hard disk. In this configuration, it can accommodate up to 26 tests with each test consisting of up to 10 content areas. Each content area can have up to 100 items in the test bank. COGENT can manage the statistical data for up to 31 classes, each class consisting of up to 50 students.

COGENT is currently being converted to the Zenith model-120 with a 10 megabyte hard disk. This version will be able to manage multiple courses/curriculums with increased statistical reporting capabilities. Both versions utilize an Epson 80 column printer for the printing of reports and tests and a Scantron model 1200 optical mark reader for the grading of the students answer sheets. The selection of hardware was based solely on existing in-house equipment.

Mircocomputer-Based Field Testing for Human Performance Assessment

P. J. Merkle, Jr., R. S. Kennedy, M. G. Smith,
J. H. Johnson (Essex Corporation)

Abstract

The chief advantages of paper-and-pencil instruments for field research are economy and simplicity, but they can have low response rates, exhibit questionable subject anonymity and security, and typically they require a proctor. During on-site investigations of the side-effects from flight simulator operations, paper-and-pencil forms and an Automated Portable Test System (APTS) were administered. The APTS functioned reliably, data production increased, and a high incidence of simulator side-effect symptomatology was detected.

The APTS is comprised of a test battery and questionnaires embodied in a microcomputer. The battery includes tests of cognition, information processing, psychomotor skill, memory, reasoning, and others. The questionnaires are: a mood adjective checklist, motion sickness symptomatology (with automated scoring), and motion sickness history. Originally conceived as a behavioral toxicology assessment tool, the APTS has applications for personnel selection and classification, education and training assessment, clinical diagnosis, and health care delivery systems.

Several military and university facilities have purchased, rented or borrowed an APTS and have begun collecting data to determine the sensitivity to various drug, environment or treatment conditions. In these preliminary studies one or more tests have been shown sensitive to morphine, chemo-radiotherapy, sleep loss, hypoxia, amphetamine, hyoscine, etc. Some of these results are reviewed.

Introduction

Prior to the advent of low cost, microcomputer systems (ca., 1972), computer technology was slow to find a place in human performance and laboratories. Several factors contributed to this lag. First and foremost, the cost of minicomputers was prohibitive for most facilities, while that of mainframe systems continues to be so. Additionally, the time-sharing operation mode of undedicated computers, required to make them cost-effective, can make them less reliable assessment instruments than traditional methods; the speed with which such computers operate is dependent upon the number of users accessing the system at any specific time. This factor results in inconsistent timing parameters from one testing session to the next. If the system is "down" for maintenance or development, all data collection comes to an abrupt halt. Although mainframes and minicomputers opened new frontiers for study, they did not prove to be cost-effective for most behavioral science laboratories.

Microcomputers expand the potential for the study of psychological phenomena in at least two ways. First, they permit more comprehensive measurement than traditional tests by providing latency and other

information not ordinarily available with traditional approaches. Second, they are capable of controlling devices which produce speech, and they are suited for complex video displays, thereby increasing the number of sensory modalities that can be involved in a testing situation. Microcomputers are also well suited for memory tests, because unlike paper-and-pencil methods, they can present stimuli for short periods of time.

Potential advantages of microcomputers for psychological testing may include: (1) identification of true deficit in performance from developmental problems; (2) standardized presentation which may lead to improved comparability of tests; (3) higher test reliabilities due to more accurate control of stimulus material; (4) bypassing of infirmities (e.g., memory deficits, dyslexia, dystrophy) of certain groups, with performance testing in innovative modes (e.g., voice recognition, touch panels, large or back-lighted keyboards, eye-tracked systems); (5) more comprehensive assessment of individuals; (6) potential for new assessment paradigms and perspectives for understanding of human performance; and (7) possible provision of more intrinsically motivating tests to subjects.

Automated Portable Test System Overview

The AUTOMATED PORTABLE TEST SYSTEM (APTS) is the first complete, compact (can be hand-held) system of its kind. It is being produced expressly for human performance assessment whether in unusual environments, with toxic substances, or with other treatments. APTS is capable of controlling and administering complex psychological testing routines, while entering and collating responses and latencies with accuracy and precision. The battery includes tests of cognition, information processing, psychomotor skill, memory, reasoning, and others.

The APTS makes it possible to maintain a professional research workspace. Any data collected or documents written on the APTS can be printed out for examination, sent via modem to the home office, or loaded onto an external cartridge for easy transport to other locations. Also, information from other locations can be communicated through external cartridge or cassette tape (new tests to use, modified data formats, etc.).

The APTS is comprised of three subsystems: (1) hardware; (2) test programs; and (3) system control.

Hardware

The hardware subsystem has been developed around a notebook-sized 8-bit personal computer: the NEC PC 8201A. Integral to the microcomputer is a 32K internal read-only memory (ROM) containing, in addition to TELCOM and TEXT EDITOR, a version of Microsoft BASIC. The technical features of the microcomputer are more fully described in NEC User's Guide (1983). Within the small, lightweight package, the system has: substantial onboard random access memory (RAM) capacity expandable to 96K; an external battery option (8 A h) providing for more than 100 h of continuous operation; and a built-in display. Refer to Figure 1 for a hardware overview.

Augmenting the notebook microcomputer are the wide variety of auxiliary components. Among these, the (32K) RAM cartridges have proved particularly

useful in applications to date. For field applications, the APT System and Testing Programs are maintained in internal RAM, and, after data collection, data are transferred to a RAM cartridge for mailing or carrying from remote sites to a centralized data-base location. (See Figure 1) For laboratory applications, it is anticipated that researchers may find it useful to extend the capabilities of the microcomputer with an external display (CRT), floppy disks, and computer interfaces. Overall, the NEC PC 8201A has the expansion options required for a wide range of field and laboratory applications.

Test Programs

The APTS component programs are developed following an iterative three-stage process: identification, mechanization, and evaluation. The identification, until now, has been on the basis of sound metric properties. Future considerations will include operational relevance and prediction and construct validity. The mechanization is conducted in-house and is the work of Essex Orlando's Chief of Systems (M. G. Smith). The programming is in BASIC (Microsoft) and Assembly language. Evaluation has subjected microbased tests to a repeated measures analysis where they were compared to their paper-and-pencil analogues. Although there have been exceptions (and these were predictable), nearly all have exhibited strong commonalities.

System Control

The APTS has been developed to provide a human assessment capability suitable for use in remote operational settings. As presented in the system overview, the hardware, test program, and system control subsystems meet the requirements for such a system. The notebook-sized NEC PC 8201A provides the basis for an easily transportable and flexible assessment system with expansion options required for a wide range of field and laboratory applications. Additionally, the development of test programs is being conducted by a process to assure efficiency and construct validity. This process is based both on evaluation tools developed for computer tests and on lessons learned during the PETER Program (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984; Smith, Krause, Kennedy, Bittner, & Harbeson, 1983). Lastly, the experimental control subsystem has been simplified for use by paraprofessionals with minimal training.

The APTS has substantial prospects for future growth and development. Attesting to this are recent and ongoing studies that have indicated that it has considerable promise for use in a broad range of unusual environments. For example, both an explosive decompression study and flight testing have found that the system is suitable for high-altitude, chamber, or airborne applications. These applications, coupled with the evaluation of the design for NASA and NSF, have indicated that the system could easily be adapted for orbiting shuttle or space station use by applying spray coatings to the interior and to the exterior case. In addition, the NEC PC 8201A has demonstrated robust capabilities to operate in at least the range of 0° to 32° C, to survive drop tests, and to withstand multiple airport x-ray exposures. The reliability of the system has been demonstrated during extensive field studies (>10^3 operational hours without failure).

Dr. Sam Schiflett of the USAF Aerospace Medical Research Laboratory at Brooks AFB, Texas, is using the short battery (Grammatical Reasoning, Pattern Recognition, Code Substitution, and Tapping) at two different altitudes in a parametric study. Thus far he has shown that performances are degraded in the following ways: greater deficits are seen at 25,000 ft. vs. 18,000 ft., and cognitive performances are more disrupted than motor.

Dr. Mary Williams at the University of New Orleans is studying whether learning curves of persons with identified learning disabilities reveal different slopes and different acquisition rates for different tasks, and/or whether there are relationships between initial score, rate of acquisition, and terminal score.

Dr. Darryl Mellard of the Institute for Research in Learning Disabilities at the University of Kansas, Lawrence, is working on a contract with the California Community College system to develop and establish eligibility criteria by identifying persons with learning disabilities. A sample of persons diagnosed as learning disabled are practicing 10 different tests following a similar paradigm to that being used by Dr. Mary Williams. One line of this research is the study of individual differences in the rate of acquisition of LD subjects versus normals.

Dr. Todd Jones of the US Coast Guard in Washington, D.C., is using several NEC PCs implemented with the short performance batteries to study fatigue at sea as well as the effects of ship motion on performance.

Dr. Lou Bandaret of the Army Natick Laboratories in Massachusetts is examining "Tower of Hanoi" and other complex games as tests. He is interested in performance effects of altitude and thermal stress. A study is now underway at that facility, under the direction of Dr. Charles Houston, on a protracted, simulated climb of Mt. Everest.

Dr. James May at the University of New Orleans is comparing the effects of performance of pre- and postexposure to optokinetic stimulation and pseudo-Coriolis stimulation to determine normals and persons with bilateral labyrinthine defects. He also has masking, meta contrast, and other temporally-based vision tests implemented on a NEC PC8201A, and is doing pilot work with learning disabilities.

Dr. Ann Streissguth at the University of Washington Medical Center, Seattle, is conducting a series of studies on the effects of Fetal Alcohol Syndrome on human performances. There are suggestions that recent memory loss is a key ingredient in the Fetal Alcohol Syndrome, and one test on the NEC PC8201A, the complex counting test which assesses a person's ability to keep track of several things at once with changing states, may be particularly useful.

Under contract to the Naval Training Equipment Center, Essex has tested 700-800 student pilots before and just after they were exposed to a ground-based flight trainer. In those studies > 10% reported motion sickness-like symptomatology. Performance data are being compared with subject reports.

Essex, through contracts with NASA and NSF, has evaluated the APTS and is issuing three reports. A final report of the Portable Human Assessment Battery (PHAB) has been submitted to NSF. The experimental work for these efforts has been conducted by R. L. Wilkes, at Casper College, Wyoming. For a complete reference list see Kennedy, Dunlap, Wilkes, and Lane MTA 1985.

Thus far, two studies have been completed for NASA(I and II). In the NASA I study, The proposed 6-minute (APTS) battery was administered four times along with analogous paper-and-pencil tests. Performance appeared stable and reliable. Two factors emerged. Reprints are available of this study and the preliminary work which preceded it. The NASA II study administered a longer battery and those data are presently being analyzed.

The study for NSF incorporated a long (10-minute) battery administered over 10 sessions and compared performance with measures of IQ (WAIS). Performance was stable for seven of the tests (vertical math and dynamic visual acuity [new tests] did not fare well). Four factors emerged. Good correlations with Performance Scale WAIS scores (multiple R = .89) were obtained; somewhat poorer with Verbal Scale scores (multiple R = .67). A preliminary draft is available.

Dr. Randall Kohl at NASA's Space Biomedical Research Institute, Houston, is undertaking a repeated measures performance testing paradigm using motion sickness drugs and provocative motion sickness tests while performance is assessed. Data collection is underway.

Dr. Pia Parth at the Fred Hutchinson Cancer Research Institute, Seattle, Washington, is administering a battery of tests from the PETER program on a NEC PC8201A to patients who have received bone marrow transplants as well as chemo-radiotherapy, and to a related cohort control group over many months of replications. The project is in its 9th month and clear-cut differences in both learning and performance are evident in the treated group. Subsidiary studies are ongoing using morphine where performance decrements on different tests were shown.

Dr. Charles Wood at Louisiana State University in Shreveport, Louisiana, is studying the effects of dexedrine, hyoscine, and scopolamine on performance on the short (APTS) battery. Preliminary findings on eight subjects show performances in predicted directions with some statistically significant. Longer tests and more subjects are planned for future studies.

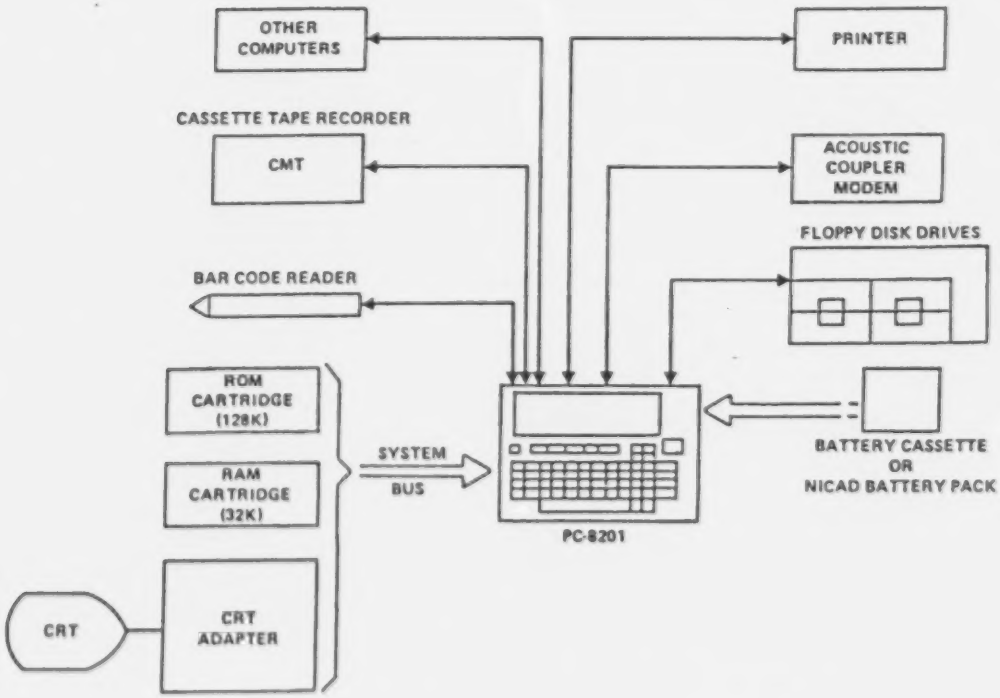
CAPT Wayne Coussens at the Tripler Hospital in Hawaii is in the process of evaluating the effects of work at altitude and mountain sickness on performance.

CDR Charles Hutchins and his students at the US Naval Postgraduate School, Monterey, California, has shown that up to 40 hours sleep loss reduces performance ($p < .02$) on APTS tests, particularly Code Substitution.

Dr. Michael McCauley of Monterey Technology, Inc., Monterey, California, under a US Coast Guard contract, has collected some at-sea data to study workload, fatigue, and environmental stress.

The APTS has proven to be rugged and reliable, but it is the first generation of such a device. As more test are implemented on the current APTS and as new computer technologies develop an even more flexable system will emerge. A current limitation of the APTS is the Liquid Crystal Display (LCD) which is difficult to read under some lighting conditions. It is hoped that new flat panel displays will be integrated into the next generation of portable computers and hence a future generation of the APTS. Future plans have also been made to transport the APTS to the IBM PC compatable environment. This transportation will broaden the scope of the APTS applications.

Support for this project was provided by National Aeronautics and Space Administration Contract NAS 9-16982 and the National Science Foundation, Contract No. 00559



OVERVIEW OF PERIPHERAL INTERFACES

FIGURE 1

Applications of a Portable Training/Testing Computer

Joan Harman, Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences

In recent years, the U.S. Army has developed and implemented increasingly complex and sophisticated weapons and communications systems. During this same period, entry-level soldiers have demonstrated declining reading, writing and computing skills. The discontinuity between increasingly demanding jobs and decreasingly skilled personnel has constituted a substantial training problem for the Army. One means of addressing this problem has been provided by the U.S. Army Research Institute (ARI) in the form of a hand-held, portable, computerized tutor that teaches technical, job-related subject matter.

Most computer-based instructional systems consist of desk top devices that are very costly and are confined to a site to which users are also confined in order to benefit from the instruction provided. The four-pound, battery-operated device called the Hand-Held Tutor, that was developed in accordance with ARI's specifications, is intended for out-of-classroom environments (mess halls, motorpools, barracks, etc.) to convert waiting periods into training opportunities. Each soldier, therefore can work with a tutor independently in a variety of settings rather than sharing a microcomputer terminal at a fixed location. The device was also required to incorporate the following features:

1- diagnostic pretests

2- self-paced instruction

3- gaming

4- instruction compatible with varying initial knowledge levels
initial motivation levels
rates of learning

5- frequent corrective feedback

The exterior features of the tutor include a 9" by 11" plastic case with an indentation molded on its top surface to hold a 5" by 5" booklet that provides instructional information and directions for interacting with the computer. Above the booklet is a multifunction liquid crystal

diode display screen that includes a two digit counter and twenty-nine character space for questions, instructions, definitions and feedback. Below the booklet is a keyboard equipped with domed conductors to provide tactile feedback to the user. The keyboard displays numerals 0 through 9, letters A through E, and the words SAY, ERASE, and GO. Beside the display screen on the upper front surface of the tutor is a built-in speaker and in the rear, a jack for alternative earphones. Also on the back of the tutor casing is a jack for a battery recharger, a switch/volume control, and a receptacle for plug-in modules that encase a computer chip programmed for the Military Occupational Specialty (MOS) instruction provided in the accompanying courseware booklet. The plug-in nature of the module offers the potential to permit the essential hardware features to accommodate a great variety of MOS instruction. Selections of all features of the tutor were based on cost, availability and human factors considerations. For example, the display screen was chosen to optimize brightness, contrast ratio, size, character font and legibility within size and cost constraints. The printed courseware booklet represents an economical alternative to systems that store text and graphics in computer memory for display on a CRT.

The major considerations in courseware development included multiple teaching techniques (gaming, drill and practice, etc.) to maximize a match with individual learning styles, initial knowledge levels and rates of learning. Users can make selections from a menu of teaching/testing options that include gaming. The booklet includes many pictures and other graphic presentations and the computer provides both immediate and delayed visual and oral feedback to responses to multiple choice questions.

The courseware is divided into units that are sequenced from less to more difficult to promote an early experience of success by the user. Each unit consists of a Pretest, Explanation, Picture Battle and Word War. Users can choose any unit to work with and any component within the unit selected.

The Pretests are short tests that are intended to establish whether the user is knowledgable about the subject matter being presented. If all but one or every question is answered correctly, the final score is presented vocally and the user is permitted to move to any other component or any other unit or, if desired, to review the Pretest. If more than one answer is wrong, the user is directed to return to the first Pretest item, reviews the test with accompanying corrective feedback, and then is directed to the Explanation component in which the subject matter is taught. This component includes test questions as a check on the progress of the instruction.

The Picture Battle component requires matching pictures or graphic presentations with visual/oral stimuli. This component displays projectiles at each end of the display screen representing friendly and enemy targets. Correct responses result in movement of the friendly projectile toward the enemy target and incorrect responses result in the same kind of movement of the enemy projectile. The objective is to destroy the enemy

target before it reaches the friendly one. The impact with the enemy target is accompanied by a sound resembling an artillery shell exploding. The impact with the friendly target only results in both projectiles returning to staring positions to re-start the game.

Word War is a component that is independent of the booklet. Both questions and multiple choice answers are presented by the computer in the form of electronic flashcards on the display screen. The instructional method calls for drill and practice in an increasing ratio review format. That is, incorrect responses result in the question being presented again after one succeeding question, and once again after three additional items have been presented. Multiple choice answers to questions answered incorrectly are randomly selected from other choices stored in the tutor's chip. Also, the position of the correct answer choice is randomly varied. The success of increasing ratio review has been demonstrated to shift learned information from short to long term memory.

The tutor, therefore, incorporates varying teaching techniques, presentation modes and kinds of feedback in order to enhance acquisition and retention of the selected subject matter. The courseware is heavily weighted with frequent, short tests to permit the user to monitor progress in acquiring the needed information and to focus attention on the most relevant materials.

In 1981, ARI awarded a contract to Franklin Research Center (FRC) and its subcontractor, Educational Testing Service (ETS) for initial development of the tutor. To test the feasibility of the device, it was decided to take advantage of the existing research foundation in computer-based methods for training vocabulary.

Accordingly, a vocabulary tutor was developed to teach technical terminology to Cannon Crewmen. This tutor was evaluated at Fort Polk, Louisiana, and Fort Drum, New York. Results demonstrated substantial increases in scores on a vocabulary test, that soldiers enjoyed using the tutor and that they found it trouble-free and easy to use. Most soldiers tried all of the components in the units and most completed all of the units. There was a considerable difference between the fastest and slowest time required to complete all units, which suggests that self-pacing is an appropriate feature of the program.

In 1983 FRC and ETS were awarded a contract that called for the adaptation of the tutor to teach mathematics. The courseware was developed to address the needs of MOS 12B, Combat Engineeers. In addition, the contractor constructed a RS-232-C serial interface for the tutor. This effort provided a hardware/software data link through which the tutor can communicate with other computers. The demonstration model permits the desk top microcomputer to download course materials to the tutor, which can then be disconnected and transported to another site for study. A diagnostic feature allows for the microcomputer to upload responses to test questions, assess the needs of the user, then download appropriate homework on which the user can practice before returning for retesting.

This development greatly increases the flexibility of the tutor and provides the potential for storing instructional materials for a variety of MOS, each set of which can be transferred to the portable device as needed.

In 1984, ETS and its subcontractor, Advanced Technology Laboratories, were awarded a contract that required development of plug-in modules to accompany the mathematics courseware. This application of the tutor is under evaluation now at the Naval Ordnance Station in Indian Head, Maryland. This evaluation site is appropriate because the mathematics needs of the MOS 12B soldiers proved to be of so general a nature that a broad range of service members are expected to benefit from the instruction. In addition, the contractor is in the process of adapting the tutor to provide M1 Tank Commanders with instruction in fire commands and degraded mode gunnery. This application will be evaluated in January of 1986.

The next application of the tutor planned by ARI is for instruction in English-as-a-second language. We intend that this version incorporate a miniaturized tape recorder to simultaneously teach reading and understanding spoken English as well as pronunciation.

DESIGN OF AN OCCUPATIONAL DATA ANALYSIS SYSTEM

MAJOR C.P. WHEELER ROYAL ARMY EDUCATIONAL CORPS

ARMY SCHOOL OF TRAINING SUPPORT

INTRODUCTION

1.The British Army has for some time based its occupational data analysis programme on a version of CODAP implemented in the IBM format and currently running on an IBM 3083 mainframe computer. This version has been superseded by CODAP 80 which was considered to be the natural progression in CODAP evolution. Its proposed implementation however highlighted difficulties that would constrain the exploitation of the power of CODAP 80. It was decided therefore, that in view of the progress made by the United States Air Force Human Resources Laboratory in the development of Advanced CODAP, the time was appropriate to review the Army`s requirements and investigate the possibility of creating a dedicated occupational data analysis system.

AIM

2.The aim of this paper is to describe the analysis of the total requirement of the Army and the subsequent design and specification of a suitable system that best meets those requirements.

SYSTEM ANALYSIS

3.The decision to use a formalised system analysis technique was taken at an early stage in the investigation. The technique chosen was Learmonth and Burchett`s Structured System Analysis and Design Method, (SSADM) which is a Ministry of Defence approved system. SSADM is a data driven method which takes as input an initial statement of requirement and produces the following outputs:

a. Program specifications
b. User clerical procedures

c. Operating instructions
d. File design or data base schema
e. Plan for testing and quality assurance

4.System analysis and design is tackled as six phases with each phase broken into steps and activities. There are clearly defined interfaces between steps in the form of working documents and criteria for review and project reporting. These six phases fall nicely into two sections:

a. The What - Systems Analysis
   (i). Analysis of Current System
   (ii). Specification of the Required System
   (iii). Selection of an Option for Implementation

b. The How - Systems Design
   (i). Detailed Data Flow Design
   (ii). Detailed Procedure/Processing Design
   (iii). Optimisation of the Physical Design

Current System Analysis

5. The system analysis in this instance was a fairly simple process and did

not require a great depth of investigation of data flow. The current system, though incorporating an automatic data processing procedure, is fundamentally a manual one. There is no standardised questionnaire format and the design is very much a matter of study team style. The transcription of raw data from the questionnaire to a computer readable format is a data preparation bureau activity which produces a magnetic tape containing the data in ICL format. This tape is mailed to the computer centre together with a magnetic tape transcription of the job control cards and run within the CODAP shell. Results printouts are sent to the initiating agency for subsequent analysis, and, if necessary, re-runs of the data with other options from the CODAP suite of programs are completed. Finally a report is produced with appropriate recommendations. A modified data flow diagram is given at figure 1 which, for the purposes of this paper, adequately describes the current system.

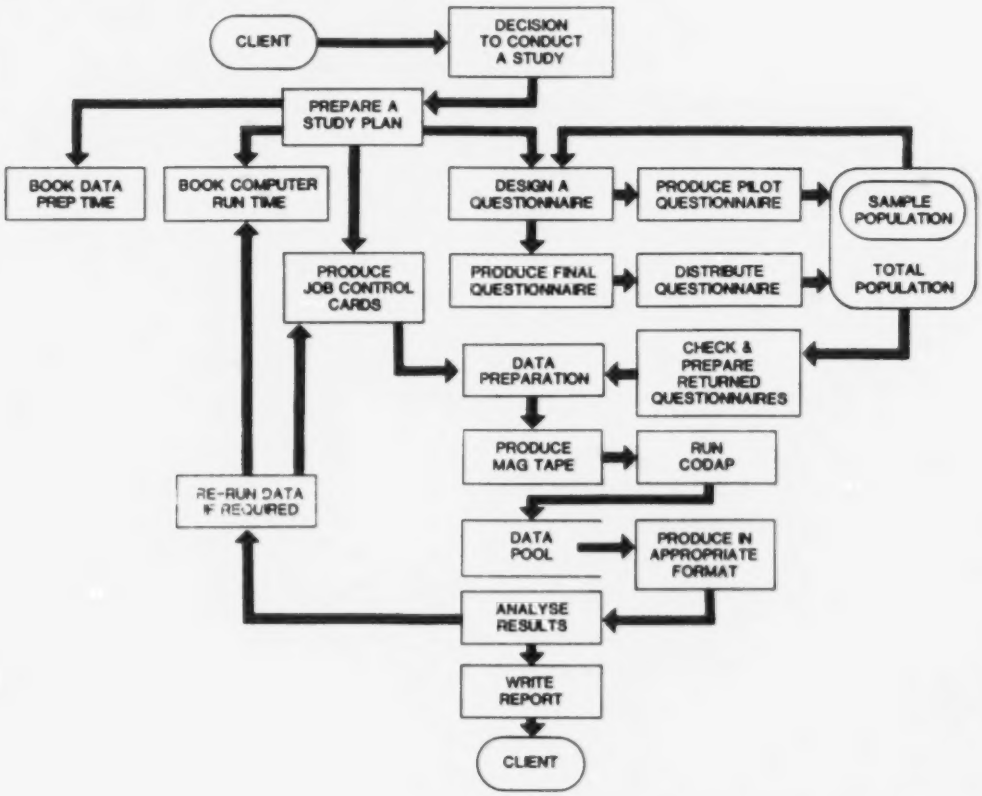


figure 1 - Current System Data Flow Diagram

The production of this document achieved two aims, firstly it ensured the analyst had a complete understanding of the system and secondly, it enabled him to identify those processes that were responsible for system operational shortfalls.

Problem Definition

6. There were two areas where problems occured, those that were inherent in the CODAP system and those that were a result of the particular implementation

at the Royal Army Pay Corps Computer Centre,(RAPC CC).They are addressed separately below.

a. Inherent Problems

(i). The version of CODAP in use is a translation of the original USAF SPERRY 1100 CODAP and reflects the state of development in 1977, (the year of acquisition). On-line data base interrogation is not possible. Confirmation of a point or the establishment of a trend requires a re-run of the raw data with fresh job control cards. This process is lengthy as fresh computer run-time slots need to be booked.

(ii). The processed data is output in the form of tables which require some expertise and patience to translate. When identifying trends this is arduous and encourages misinterpretation of the facts.

(iii). There is a limit of 999 task history and/or time data items available to each study. Though for some studies this is more than adequate in many of those proposed it may well be a constraining factor. In a recent study of Army majors` education the questionnaire had to be "sliced" to accommodate some 2200 data items from 600 respondents. This considerably added to the analysis task.

b. Implementation Problems -

(i). CODAP was designed to utilise Optical Mark Reader, (OMR) data input peripherals. The host Computer Centre did not permit the use of this facility in the original implementation for security reasons. It was feared ASTS control of the data,(and by implication the computer), would compromise the integrity of the system and its other applications programs. The result is a resource intense and time consuming data preparation process. Currently based on a bureau, the data magnetic tape is prepared in ICL format which has to be transcribed into IBM format prior to run-time. Slippages do occur in this procedure giving worst case times in the order of 25 weeks for data preparation. The best case was 3 weeks with an average of 10 weeks.

(ii). CODAP is a low priority task at the Computer Centre consequently their resource and time allocation to it is limited. This further aggravates the problem indicated above.

Specification of the Required System

7.Traditional job analyses for training have concentrated on the collection, amongst others, of information relating to task difficulty, importance and frequency, (DIF analysis). Though apparently simplistic, all of these factors, if taken in isolation are liable to produce data that will ultimately result in a wrong interpretation by the analyst and subsequent incorrect training decisions being made.

8.To seek estimates of task difficulty alone can result in invalid training decisions. Estimates can be made by the job incumbent based on differing factors, size, weight, environment, availability of spares as well as the more obvious and expected responses concerning lack of training, lack of practise or lack of ability. It is now accepted that a less ambiguous measure of difficulty is the time taken to learn to perform the task to some required level. This factor has some value in skill retention analysis.

9.Similarly estimates of importance of a task are of little value without further amplification of the frame of reference. The estimates of "importance to whom" and "importance for what" can become quite subjective and difficult to quantify. If it is importance that we wish to measure then the job incumbent´s supervising officer is better placed to make that assessment. Within defence requirements, importance can be considered to be a composite of:

a. Immediacy of skill requirement after training

b. Consequence of inadequate training

10.Precise measures of frequency of performance cannot be made by job incumbents, often due to the cyclic or seasonal nature of jobs. They can, however, indicate that they spend more time on one task than another in relative terms. It is accepted that such a relative time scale can produce a more reliable measure of time spent on tasks and that subsequent calculation of percentage involvement is more accurate.

11.These factors were accepted by the Army at the time CODAP was first implemented, in reality it was a major reason for its aquisition. The situation has not changed, CODAP is still considered to be the most appropriate analytical tool available to military occupational analysts and training designers. It was decided that any future occupational data analysis system should be based on the most recent development of CODAP. Therefore system design should concentrate on improving those areas of operational shortfall identified in the analysis.

12. The most important requirement for a modern occupational data analysis system is an automated data capture facility. Data can be rapidly transcribed from standardised questionnaires and input directly to computer memory. This is an OMR operation. The ability to interrogate the data base in real-time is an advantage. The analyst should be able to manipulate data in an interactive manner using an on-line query language. Care also must be taken to ensure the system file and record lengths are of sufficient magnitude to handle the largest studies. Failing this ´slicing´ of the data should not seriously detriment the study.

13.The output of the system should be readable with an appropriate use of graphics to display data in a format such that trends are highlighted. A more thorough examination of data relevant to the problem is then possible. The total system must be able to respond rapidly, (within one week), to raw data input from respondents. It must therefore be co-located with the system managers, the system hardware being dedicated to occupational data analysis.

Possible Solutions

14. The alternatives available to ensure the continuation of an occupational data analysis service are as follows:

Option 1 - Continue with the existing system.
Option 2 - Implement CODAP 80 on the existing computer.
Option 3 - Implement a commercial package, (eg. SPSS), at the RAEC Centre.
Option 4 - Implement USAF CODAP at the RAEC Centre.

15.Option 1 - Though this is a solution option, the problems outlined previously preclude it from consideration in all but the most extreme of circumstances. The service available to the Army's Training Organisation would be subject to continual degradation leading eventually to inefficient training with all that that implies.

16.Option 2 - CODAP 80 was designed to overcome the deficiencies of CODAP. It is written in FORTRAN and well documented. There have however, been some shortfalls in the expected program run-time target efficiency. Though this may well be acceptable on a dedicated system, this is not the case with our host computer. Run-times measured in hours or even days would consign CODAP studies to those periods least used by the operational system. The outcome of this would be an even longer turn around time for processed data. It is possible that the RAPC CC would consider the task too resource demanding and not continue to host CODAP. A second and perhaps more important point is the provision of interactive facilities to ASTS. Users of CODAP 80 are able to interrogate the data base on-line. The analyst is able therefore, to manipulate the data and run the requisite program that best meets his needs from a remote terminal. To achieve this a communications link would need to be established between ASTS and RAPC CC, (a distance of some 70 miles). The link would be a British Telecom leased line, (telephone line), and would require IBM compatible terminals at each end to ensure correct communications protocol. At this time however, the RAPC CC insist on the terminal being able only to access data and not to manipulate it. The reason for this is the security of the RAPC CC computer system may be compromised and unauthorised access to other programs possible. CODAP 80 is essentially an interactive medium and without the ability to sort data and run programs from a local terminal, much of its power would be lost. Similarly CODAP 80 can accept data from OMR data input devices, in the past RAPC CC have not permitted their use.

17.Option 3 - Several commercial packages were investigated and a general comment on their suitability was that none were specifically designed for occupational data analysis in a military environment. Also as far as can ascertained from the literature they are not able to produce a job description. Commercial packages are quite expensive , typical costs over a five year period are £47750, (costs for SPSS).

18. Option 4 - This option is based on the most recent implementation of the USAF CODAP running on a SPERRY UNIVAC 1100/81 mainframe computer. A preliminary investigation into this system is encouraging, it would appear that those shortfalls of the current CODAP system have been overcome. A representative of ASTS will shortly visit the USAF HRL at Brooks AFB to confirm the suitability of it to meet the requirements of the British Army. An implementation of this version of CODAP would be hosted on a SPERRY Series 11 computer running under the SPERRY 110C Operating System and located at ASTS.

Proposed System

19.The proposed system should be based on Option 4 and should consist of USAF CODAP, (if found to be suitable), running on a SPERRY System 11 computer with automated data input peripherals, OMR, and appropriate output devices, VDUs and printers. This system to be located at ASTS and managed by the Systems Group of the RAEC and be dedicated to occupational data analysis. The departure from the systems analysis routine in specifying hardware without a design phase is justified by the fact that CODAP was written for this series

of computers and exploits much of its hard wired power. Any re-configuration to utilise other computers would result in an expensive and lengthy process with no guarantee of operational efficiency at the end, thus narrowing the field of choice. The OMR question has been thoroughly investigated and a suitable machine identified. This is a KAISER OMR 80 which has the ability to read both sides of a document on each pass and identify those documents that contain errors and segregate them. The OMR 80 can read up to 10,000 documents per hour.

20.The disadvantage of this proposal is its high initial cost, in the region of £250,000. The advantages however would more than compensate for this in specific cost savings of training time and in the establishment of an effective and efficient training system. This system would be able to take advantage of future development by the USAF HRL with the improvements to the service that that entails. Also the Series 11 computer would possess some spare capacity for other tasks, eg office management systems, library and course administration which will add to the general efficiency of this establishment.

Justification

21.In control systems theory it is a fact that a system without a clear purpose and suitable controlling feed-back will eventually degenerate to chaos. The Army's Training System is based on a scientific methodology, SAT, that requires accurate training objectives derived from comprehensive job descriptions. Feed-back in the form of internal and external validation provides the control and ensures the training objectives are valid such that training courses fit the soldier for his task. The current implementation of CODAP has short comings in the capture and presentation of data. There is some doubt as to the willingness of RAPC CC to continue to host CODAP. The time is right to make good the implementation problems with a modern version based on a dedicated computer under the direct control of ASTS.

22.It is the recommendation of this paper therefore, that:

a.With the proviso above, the USAF version of CODAP be adopted for use by the British Army.

b.A suitable purchase of hardware be made that is able to run CODAP, eg a SPERRY System 11 computer, peripherals and a KAISER OMR 80.

c.A CODAP consultancy and management cell be established at ASTS to control, advise and manage the occupational data analysis requirements of the Army with a possible extension to the Royal Navy and the Royal Air Force.

IMPLEMENTATION OF AIR FORCE ASCII CODAP: NEW VERSUS OLD SYSTEM

William J. Phalen
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

Michael R. Staley
Johnny J. Weissmuller
Texas MAXIMA Corporation
San Antonio, Texas 78209

I. NEED FOR REDESIGN OF THE CODAP SYSTEM

A. BACKGROUND

1. Introduction

The principal occupational analysis technology in the Air Force is the Comprehensive Occupational Data Analysis Programs (CODAP) software system, which has supported a major occupational research program within the Air Force Human Resources Laboratory (AFHRL) and a major operational occupational analysis program within the Air Force Occupational Measurement Center (USAFOMC) since 1967. From a software standpoint, CODAP can be defined as a package of computer programs used to input, process, organize, and report occupational data from job inventories. From a measurement standpoint, CODAP can be defined as a set of procedures which focus on the analysis and comparison of individual and group job descriptions and their associated biographical data, as well as on individual tasks and groups of tasks (task modules) and their associated characteristics. From an applications standpoint, CODAP can be defined in terms of its significant contributions to the mission accomplishment of the Air Force manpower, personnel, and training management (MPT) functions. These contributions include: providing a data-based approach to evaluating and updating Air Force officer and enlisted classification structures, providing an empirical means of restructuring and redesigning jobs, providing data that has been instrumental in eliminating unnecessary training and in pinpointing specific training requirements, and providing a scientifically sound basis for realigning entry-level aptitude requirements across career fields.

2. Problem

The CODAP system began 20 years ago as a software package of approximately 15 general-purpose programs. However, in order to keep pace with the rapidly expanding needs of occupational researchers and analysts since that time, the system was forced to expand unsystematically into a somewhat confusing aggregate of more than 60 general-purpose programs. Over time, the system became increasingly difficult to maintain, modify, or augment without extensive programmer training and experience. Many of the programs had been hastily developed to meet short suspenses, with little time available for producing proper program documentation; and many of these programs were partially redundant with already existing programs. Files maintained in the system had been put in a variety of formats and stored on a variety of media by a succession of programmers using a variety of programming standards and styles. New and advanced developments that had been needed in the system for some time required that many of the existing

programs run more efficiently and interface more easily.

The CODAP system had not only grown in size and complexity since its inception, but the evolutionary process had become increasingly more dynamic as a direct result of significant hardware improvements to the AFHRL Sperry system, major enhancements to the basic structure of the CODAP software system, a growing sophistication in the methods and procedures required by Air Force occupational researchers and analysts, and a wider range of users and applications. Consequently, it became increasingly urgent that a major system redesign effort be launched to consolidate and use to better advantage the many additions and modifications that had been incorporated into CODAP over the years and to create new technology and software to meet the current and anticipated methodological and applications requirements of CODAP users.

B. OBJECTIVES

The objectives of the CODAP redesign project were to improve Air Force occupational analysis methodology, translate methodological improvements into R&D technological capabilities, standardize occupational analysis procedures, and generate efficiencies in the occupational analysis process that would significantly improve quality of product with substantially less expenditure of scientist/analyst man-hours and computer processing time. Increased data processing efficiency would also make feasible the development of complex analysis programs, especially in the area of automated job analysis, that would otherwise have been too costly to run.

C. MEANS

Early in 1983, AFHRL negotiated a task ordering contract with an 8A (small business, minority-owned) contractor, the MAXIMA Corporation. The first task submitted to MAXIMA was a 26-month project (later extended to 29 months) to rewrite/convert the CODAP system, as needed, to bring it in line with the most recent standards for software development, some of which represented requirements established by the AFHRL Technical Services Division (TS). The initial meeting with the contractor identified the following as the most urgent needs for revision of the CODAP system:

1. Converting the general-purpose CODAP programs from FIELDATA FORTRAN, which was no longer being supported by TS, to the FORTRAN 77 Standard (ASCII FORTRAN). "ASCII" is an acronym for the American Standard Code for Information Interchange.

2. Converting the utility programs from the old, TS-developed PILOT language, which was no longer being supported by Sperry, to the new, TS-developed PRISM language.

3. Converting the system files from a variety of formats to standardized mass storage files, wherever feasible.

4. Developing simplified processing procedures (runstream generators) for often-used program strings.

5. Improving formats of printed reports to enhance readability and interpretability.

6. Conducting exploratory development of new analytic capabilities, especially in the areas of profile analysis, nonhierarchical clustering,

two-way clustering (cases x tasks), module technology, and automated job typing.

At this meeting, also, major new Air Force manpower, personnel, and training (MPT) programs which would require the extensive use of occupational survey data were discussed in terms of how these projects might impact the planning of the CODAP redesign. Examples of these emerging applications included such research and development (R&D) projects as: the Training Decisions System (TDS), the Basic Cognitive Skills Project, the Advanced On-the-Job Training System (AOTS), the Task Qualification Assessment (TQA) Project, the Task Identification and Evaluation System (TIES), and various Performance Measurement projects. In addition to major programs, the new CODAP system would be required to support new operational applications for Air Force managers and decision makers. An example of such an application would be the use of CODAP-based occupational survey data to help establish test outline "testing importance" specifications for the development of enlisted promotion tests.

As a consequence of this meeting, three major project objectives were identified: increased operational efficiency, improved system maintainability, and expanded analytic capability. Each major project objective was, in turn, further broken down into technical goals, which will be presented and discussed in the next section, no longer as goals, but as redesign accomplishments. These accomplishments show how the redesigned CODAP system (ASCII CODAP) differs from and is an improvement upon the old CODAP system (FIELDATA CODAP).

II. IMPLEMENTATION OF CODAP REDESIGN: NEW VERSUS OLD SYSTEM

As compared to the old FIELDATA CODAP system, the new ASCII CODAP system has achieved the three major project objectives listed in the previous section, as follows:

A. In ASCII CODAP, Operational Efficiency Has Been Increased.

1. Run setups have been simplified.
   - Functional program names used
   - Mnemonics employed to name control card options
   - Control cards standardized
   - Program execution by processor control cards
     - increased flexibility in filename specification
     - elimination of separate filename cards

2. Audit trails have been improved. The first page of every report contains information such as the report ID, when the report was created, the IDs of all input files, options selected, and applicable selection or cutoff parameters for cases and/or tasks.

3. Resource utilization has been improved.
   - Reduced core requirements for many programs
   - Reduced running times for many programs (e.g., job description program runs 10 times faster)
   - Elimination of three-reel files

4. Turnaround has been speeded up.
   - Reduced core requirements and running times of many frequently run, multiple-execution programs, as well as the elimination of three-reel files, has permitted daytime running of these programs.
   - Conversion of files to mass storage format has speeded up processing and allows some programs which were previously run only in batch mode to be run in demand mode.

5. Fewer runs are required to accomplish a standard analysis.
   - Elimination of redundant programs
   - Combining of programs which perform related functions
   - Development of functionally pure programs (e.g., sample selection extracted from multiple programs and consolidated into a single program)

6. Training requirements for computer technicians have been simplified.
   - Development of computer-based training package (in progress)
   - Better documentation
   - Reduced number of programs
   - Functionally pure programs
   - Mnemonic control card options
   - Visible cue report format specifications
   - Standardized control cards
   - Automatic process generation
   - Separate report file process

B. In ASCII CODAP, system maintainability has been improved.

1. Structured programming used.
2. Conversion to FORTRAN 77 Standard (ASCII FORTRAN) and PRISM.
3. Fewer programs (reduced from 120 to 83).
4. Functionally discrete programs.
5. Reduced amount of source code.
   - Assembly language code reduced
   - Total lines of code reduced (ASCII: 27,186 vs. FIELDATA: 48,477)
     - elimination of redundant code
     - interfacing with non-CODAP software rather than retaining similar CODAP programs

6. Increased internal documentation (43.4% more lines of comments).

7. Increased supporting documentation, which is all in automated form.
   - Users manual
   - Standards documentation
   - Programmer reference guide
   - Subroutine documentation
   - Subprogram documentation
   - File format documentation

8. System has been standardized, wherever feasible.
   - Program code
   - Program names
   - Program documentation
   - File formats
   - Subroutine library

9. Formalized test and acceptance procedures to ensure reliability.

C. In ASCII CODAP, analytic capability has been expanded.

1. Increased system limits.
   - From 1,700 to 3,000 tasks
   - From 1,000 to 9,999 task modules
   - From 1,726 to 3,000 tasks per module
   - From 999 to 2,000 history variables
   - From 500 to 9,999 computed variables

   Unchanged system limits.
   - 20,000 cases total
   - 7,000 cases for clustering
   - 26 duties
   - 66 characters for variable description

2. New features and capabilities.
   - Clustering technology expanded
     - additional overlap indices incorporated (four new indices)
     - task clustering capability added
     - nonhierarchical clustering capability added
     - profile analysis capability added
   - Module technology expanded
     - tasks can be summarized into modules
     - modules can be used like tasks
     - module values can be clustered
     - module data collected from job incumbents/expert raters can be processed directly
   - Job typing technology expanded
     - pairwise group comparison program (AUTOJT) capabilities expanded
     - core task analysis program (CORSET) capabilities expanded
     - program developed which makes initial selection of job types automatically (JOBSET)
   - Interrater reliability technology expanded

3. Improved final products.
   - More readable and interpretable report formats
   - Greater availability and more flexible use of free text
   - Use of upper- and lower-case letters and overstriking
   - Greater standardization of formatting for reporting similar kinds of data

4. External system interfaces to non-CODAP software.
   - Data distribution routines
   - Correlation and regression software
   - Factor analysis software
   - Other statistical packages, such as SPSS, BMDP, and IMSL

III. ISSUES RELATED TO CODAP TECHNOLOGY TRANSFER

A. Transfer of Air Force ASCII CODAP

Air Force Regulation (AFR) 300-6 governs the release of Air Force-owned or developed computer software packages. It states in para 11-71(2): "Computer programs and related technical data are not considered 'records' within the congressional intent of 5 U.S.C. 552; these items are considered property." It further states in para 11-7b(1): "Software packages will not be released to the private sector except when in the best interest of the government." In the event that Air Force software, such as ASCII CODAP, is released to a private sector firm, the requester must certify that the package will not be published for profit or in any manner offered for sale

to the government and will not be sold or given to any other activity or firm without the prior written approval of the Air Force. On the other hand, AFR 300-6 permits the release of ASCII CODAP to the public sector, such as government agencies and universities, provided that the recipient signs a statement of terms and conditions which frees the Air Force of any liability and/or responsibility for the software and requires the recipient to state the intended use of the software.

B. Cost of Compatible Hardware Versus Cost of Software Conversion

Sperry representatives were asked to speculate as to what would be the necessary configuration for a minimal Sperry system to meet the requirement that there would be 5 to 10 users on a system dedicated to running CODAP. It was agreed that a minimally adequate system would consist of a Sperry System 11 or Sperry 1100/60 or Sperry 1100/61 with a Mega Word of main memory, an 8470 disk system (and controller) containing 50K tracks (89 million words), three 1600 BPI tape drives with tape controller, console, upper-lower case printer, and a communications subsystem. Such a system would cost approximately $250,000. ASCII CODAP could be transferred to such a system at little cost and be up and running almost immediately. On the other hand, the contractors who produced ASCII CODAP estimated that it may cost as much as $250,000 and require one year for two senior systems analysts and two programmers conversant with CODAP and familiar with both the Sperry system and the non-Sperry host system to accomplish the conversion of ASCII CODAP for running on the non-Sperry system. Given these two alternatives, the choice of which way to go should be an easy one. Nevertheless, justifying the purchase of another computer system when your organization already has one is usually more difficult than getting authorization and funds to do a software conversion.

IV. CONCLUSION

The new Air Force ASCII CODAP system is now undergoing final test and acceptance by the AFHRL Technical Services Division (AFHRL/TS). The process is moving along smoothly and the new system should be ready for release to the USAF Occupational Measurement Center (USAFOMC) and other Sperry/CODAP users by 1 January 1986.

REFERENCES

AFR 300-6. (1980, 11 July). Automatic Data Processing Resource (ADPR) Management.

Phalen, W.J., Weissmuller, J.J., & Staley, M.R. (1985, May). Advanced CODAP: New analysis capabilities. Fifth International Occupational Analysts Workshop, San Antonio, TX.

Staley, M.R., Weissmuller, J.J., & Phalen, W.J. (1985, May). ASCII CODAP: The impact of system design for emerging applications. Fifth International Occupational Analysts Workshop, San Antonio, TX.

Weissmuller, J.J., Staley, M.R., & Phalen, W.J. (1985, May). ASCII CODAP: Quarterly status report. Fifth International Occupational Analysts Workshop, San Antonio, TX.

Occupational Research Data Bank (ORDB)

Martin E. Ellingsworth

Air Force Human Resources Laboratory
Brooks AFB, Texas 78235-5601

Louis F. Olivier
Glenda J. Pfeiffer

OAO Corporation
Brooks AFB, Texas 78235-5601

INTRODUCTION

The Occupational Research Data Bank (ORDB) is a computer-based occupational information system. It provides researchers and managers a means of rapid on-line retrieval of a variety of current and historical occupational information on Air Force enlisted specialties and the people performing duty in them through a set of user-friendly, tutorial programs. Use of ORDB streamlines background research, permits quick in-depth orientation to specialties, and provides a database for a rapid response capability for research and management concerns. The ORDB is of great value as a tool to increase productivity and as an instrument for longitudinal and cross-specialty analyses.

Development of the ORDB began in 1978 with the investigation of available data that would contribute to Air Force occupational analysis and management. A number of sources and types of information were identified and have been obtained for inclusion in the ORDB (Carpenter, Archer, & Camp, 1979; Stephenson, 1979; Camp, 1982). The primary contractor for this effort is the General Services Administration data processing services contractor, currently OAO Corporation. OAO personnel assigned to this project are collocated with the monitoring activity (AFHRL/MOMM) at Brooks AFB, Texas.

The ORDB is composed of both digital and hard copy data consisting of technical reports and studies, Air Force Regulation 39-1 information by career area, statistical variables summarized for occupations from individual Air Force members and technical training course data, and Comprehensive Occupational Data Analysis Programs (CODAP) (Christal, 1974) studies performed at the Air Force Occupational Measurement Center (OMC). These types of information have been obtained and are incorporated in the ORDB. The subsystems which provide for storage and on-line retrieval of the information are described in the following section. Before going into the ORDB subsystems, here are two notes of interest:

1. When referencing an Air Force Specialty Code (AFSC), there are three levels of detail that can be shown using the five numbers of the AFSC. Here are the types, abbreviations, and examples of these three levels.

Skill Level (S) e.g., 27230, 44550G (skill levels 1, 3, 5, 7, 9, & 0)
Ladder (L) e.g., 973X2, 113X0C
Career (C) e.g., 81XXX, 42XXX

2. References to the figures at the end of this paper are actual slides from the elaborative case study scenario of the 426X2 (Jet Engine Mechanic) Ladder AFSC, which was the major part of the briefing illustrating how an occupational researcher could benefit from using the ORDB. Figures 1 and 2 are actual output from the ORDB.

SYSTEM OVERVIEW

The ORDB operates on the AFHRL Sperry 1100/81. Seven subsystems are tailored to the types of data and kinds of retrieval needed by the user. These subsystems are linked together by a front-end program to simplify the use of the ORDB. The programs are designed to interact with the user, assisting in the choice of the appropriate subsystem, and in selecting the desired information. Each subsystem is described below.

1. Computer Assisted Reference Locator (CARL). The CARL subsystem is used to reference hard copy occupational data items, such as recurring reports, occupational survey reports, job inventories, films, and microfiche, which are stored at the Air Force Human Resources Laboratory, Brooks AFB, Texas. References are based on user selected keywords. CARL was obtained from the Navy Personnel Research and Development Center (NPRDC) and modified to operate on the 1100/81 (Sands, 1978; Sands & Hartman, 1979). Additional modifications were made to accept AFSCs as keywords and to clarify user selection of output options.

Each reference stored in the CARL subsystem includes such information as author, name or title of the reference, type of reference, a brief narrative description, and an associated list of keywords for each reference (see Figure 1).

To speed the referencing process, two search techniques have been added to CARL--Quick and Smart. Quick is a binary search on a given list of keywords available upon request, while Smart is a character string search across the list of available keywords. Both respond with the number of references located and ask if and how the user would like to see output. In addition, the user can expand or reduce the number of references by using additional keywords or character strings.

2. Aptitude Requirements Component (ARC). The ARC subsystem contains AFSC descriptions (for ladder and career field), progression ladders, and prerequisite data for the years 1978 to the present (1985). The ARC has an AFSC number change history file which tracks all changes from March 1965 through the present. In addition, aptitude requirements information for each AFSC is stored and accessible (Figure 2). Also, the ARC subsystem contains KINTON (a contractor) and OMC study information including Average Task Difficulty Per Unit Time Spent (ATDPUTS), validity/reliability information, statistics, and minimum/maximum task information. It should be noted that the KINTON reports (approximately 200) were produced in the Aptitude Requirements benchmarking research, while OMC studies are completed with each occupational survey of an AFSC (Kinton uses a 25-point scale while OMC uses a 9-point scale for ATDPUTS).

Since the ORDB is a menu-oriented, tutorial system, access and use of this and the following subsystems are quite easy and efficient. You only really need to have an idea about the AFSC structures in the Air Force. This structuring is readily apparent in the use of the statistics subsystem.

3. Statistical Variable. The statistics subsystem contains demographic, aptitude, education, training, turnover, and duty related information on Air Force enlisted personnel. Information is sorted by AFSC, population group, and year on a total of 125 different variables for the most current 5 years of data. Population group is based on enlistment status by Total Active Federal Military Service (TAFMS) (0-4 years, 5-8 years, 8+ years, total sample, or current year's accessions). See Figure 3 for a partial list of variables and their level of detail (L, C, or S). An AFSC must be valid for the type of detail available and it must have existed at the end of the year for which data are being requested (validity of an AFSC can be checked in the ARC subsystem, and the validity of detail can be checked on the menu available in the statistics subsystem).

While 5 years of data are stored on-line, earlier years of statistical data will be accessible via batch run. The sources are the Uniform Airman Record (UAR), Pipeline Management System (PMS), Airmen Gain and Loss (AGL), and Processing and Classification of Enlistees (PACE) files stored at AFHRL.

This subsystem uses System 2000 (S2K) Data Base Management System with COBOL extension (Intel Corporation, 1982).

4. CODAP Report Display. This subsystem was developed to provide the task scientist and manager with the ability to rapidly retrieve OMC and AFHRL CODAP reports and review them on the terminal screen. To accommodate the standard CODAP report format, Datagraphix 132 character remote terminals are in use at principal user sites. Studies can be selected by either AFSC or study number (from a menu of available studies). Studies from 1978 to the present have been loaded, and any report retrieved on the screen can also be printed at the user's option. This subsystem is programmed in the Programming Instructions for String Manipulation (PRISM) language (AFHRL/TS, 1982).

From the initial listing of CODAP reports contained in the OMC studies, a determination is made as to which CODAP reports will be loaded into the ORDB (Figure 4). The determination is based on the requirement of a report that concerns a population group and an AFSC in a similar manner to the statistics subsystem organization.

When the user selects this subsystem from the ORDB introductory screen, he or she is provided a choice of the five different CODAP retrieval features:

a. CODAP report display--The user can view the text of individual reports and obtain a hard copy of desired reports.

b. CODAP edit feature--The user can use editing commands to select lines from one or more reports from one or more studies and sort them into a customized hard copy output.

c. Task-level cross-study--The system will retrieve tasks containing key words from multiple studies and print hardcopy, if the user desires.

d. Background cross-study analysis--The system will retrieve 15 background variables from multiple studies and print hard copy, if the user desires.

e. Title decks/Specialty Training Standard (STS) items--The system will create a file containing STS items for an AFSC in the format of a title deck.

5. Cross-Study Analysis. This subsystem was developed in response to the need to compare CODAP reports across specialties. Since CODAP variable numbers and titles are not necessarily standard, identifying corresponding data in different studies presents a difficult task. To solve this problem, studies are indexed as they are loaded to the ORDB for a set of 15 variables and 8 groups. The variables include: Number of Tasks, ATDPUTS, Job Difficulty Index, Grade, Major Command, Time in Career Field (TICF), TAFMS, Eligible to Reenlist, Eligible for Retirement, Job Interest, Talent Utilization, Training Utilization, Sense of Accomplishment, Plan to Reenlist, and How Assigned to Present Career Field.

Groups that can be analyzed include: Total Sample, Skill Levels 3, 5, 7, 9, 1-48 months TAFMS or TICF, 49-96 Months TAFMS or TICF, and 97+ Months TAFMS or TICF. On-line retrieval of corresponding data from multiple studies on one or a number of job groups can be performed using this system (Figure 5). For example, job difficulty of airmen with 1-48 months TAFMS can be retrieved for comparison across any number of AFSCs. This subsystem is written in PRISM and uses the same data files as the CODAP Report Display subsystem.

6. Statistical Package for the Social Sciences (SPSS)/Statistical Interface. A total of four SPSS procedures are interfaced with the Statistical Variable Subsystem of the ORDB: ANOVA, BREAKDOWN, T-TEST, and CROSSTABS (Figure 6). ORDB allows the user to produce statistical analyses of ORDB variables using SPSS without requiring him to be familiar with formatting SPSS run cards. The interface program provides easy to follow instructions for user inputs.

The user may initiate a batch run which will automatically retrieve the ORDB statistics and create a runstream of SPSS control cards. This will result in an SPSS run with output. The user has the choice, however, of running the SPSS automatically or having the file containing the runstream retained for additional modification.

7. Comments. The comments subsystem provides an opportunity for users and developers to record information related to the ORDB while using a remote terminal. Comments can include anything relevant to data contained in the system, or to the system operation itself. It has been especially useful as a means of obtaining user feedback and for announcing the implementation of enhancements or changes.

CONCLUSION

ORDB is currently used by many of the major projects at AFHRL (e.g., Training Decision Systems, Basic Skills, Aptitude Requirements, Performance Measurement, and RIVET work force, and it is planned for use with Advanced On-the-Job Training). At OMC, ORDB is used to provide quick in-depth orientations to AFSCs as well as to rapidly respond to high level management queries.

ORDB relates many dispersed sets of data into a consolidated data bank which can be rapidly accessed. Instead of the normal laborious and time-consuming task of finding background information by formal requests to computer databases, searching Air Force regulations, and/or digging through a library of technical reports and previous studies, the ORDB allows the user to streamline data retrieval while saving computer resources. ORDB is valuable for aiding research design, for conducting historical and cross-specialty analyses, and for guarding against duplication of effort and inconsistencies between databases. Clearly, ORDB enhances researcher and management productivity.

References

AFHRL/TS. (1982). Programming instructions for string manipulation (PRISM). Developed and maintained by Air Force Human Resources Laboratory, Technical Services Division, Brooks AFB, TX.

Camp, R.L. (1982, October). Implementation of an Air Force occupational research data bank. Proceedings of the 24th Annual Conference of the Military Testing Association, San Antonio, TX.

Carpenter, J.B., Archer, W.B., & Camp, R.L. (1979, October). Establishing an Air Force occupational research data bank. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, CA.

Christal, R.E. The United States Air Force occupational research project. (1974, January). AFHRL-TR-73-75, AD-774-574. Occupational Research Division, Air Force Human Resources Laboratory, Lackland AFB, TX.

Intel Corporation. (1982). System 2000 (S2K) data base management system. Intel Corporation, P.O. Box 9968, Austin, TX.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. (1975). Statistical package for the Social Sciences (SPSS) (2nd ed.). McGraw-Hill, St. Louis.

Sands, W.A. (1978, October). Computer assisted reference locator (CARL) system: An overview. Proceedings of the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK.

Sands, W.A., & Hartman, L. (1979, October). Research management applications of the computer assisted reference locator (CARL) system. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, CA.

Stephenson, R.W. (1979, October). Development plans for an Air Force occupational research data bank. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, CA.

REFERENCE 1436

OMC-CODAP 80J

AFSC 426X2 JET ENGINE/426X3 TURBOPROP PROPULSION (MULTILADDER STUDY).

AFPT# 90-426-424. S# 7662. JI DATE OCT 81 RCVD. DUTIES=16. TASKS=587.

DEVELOPER-DUBOIS. OSR DATE APR 82 RCVD. ANALYST-ALTON. FICHE-FOUR MICROFICHE.

426X2 REGULAR ANALYSIS RCVD; 426X2 SPECIAL F-15/16 AND TRAINING EXTRACT RCVD;

426X3 REGULAR ANALYSIS RCVD; 426X3 TRAINING EXTRACT RCVD. DATA LOCATION-STUDY OPEN.

OMC-CODAP

JI 1980	PROPULSION SYSTEMSOSR 1982		F-15 AIRCRAFT
DL-OPEN(OMC)	426X2	426X3	426XX
42XXX	AIRCRAFT SYSTEMS MAINTENANCE		F-16 AIRCRAFT

FIGURE 1

LADDER DESCRIPTION

JET ENGINE MECHANIC

AFSC: 426X2 YEAR: 1984

INSPECTS, REMOVES, INSTALLS, DISASSEMBLES, TROUBLESHOOTS, REPAIRS, ASSEMBLES SERVICES, TESTS, AND MODIFIES TURBOJET AND TURBOFAN AIRCRAFT ENGINES, TURBOJET MISSILE ENGINES, AND SMALL GAS TURBINE ENGINES.

••AFSC HISTORY••

AFSC CHANGES - 426X2	7603	CREATED FROM 43230
	7910	SOME PERSONNEL CONVERTED TO 426X3
	8304	SOME PERSONNEL CONVERTED TO 426X4

PREREQUISITES

JET ENGINE MECHANIC

AFSC: 426X2 YEAR: 1984

APTITUDE SCORES	MM4
PHYSICAL PROFILE	333133
AUTHORIZED FOR ENLISTED WOMEN	Y
PHYSICAL WORK CAPACITY	2
CERTIFICATION-LICENSE REQUIREMENTS	
MANDATORY TRAINING COURSES	7-LEVEL MANAGEMENT COURSE
INPUT FROM AFSC	99000
STS CENTER	CHANUTE TTC
OTHER PREREQUISITES	U.S. CITIZEN
	COLOR VISION REQUIRED

FIGURE 2

PARTIAL VARIABLE MENU

SEQ NO.	VARIABLE NAME	TYPE		
1	ACAD-EDUC LEVEL	L	C	
80	NCO PROF MIL EDUC	L		
22	ASVAB-ADMIN	L	C	S
47	BASE OF ASSIGNMENT	L		
67	MAJCOM ASGMT AREA	L	C	
96	AVIATION STATUS	L		
114	SEI (SPECIAL EXPERIENCE IDENTIFIER)	L		
17	AFSC-WARSKILL	L		S
60	LOSS RATE PERCENT	L	C	
71	MARITAL STATUS	L		
112	RACE	L	C	
115	SEX	L	C	

FIGURE 3

CODAP REPORT DISPLAY

- JOB SPECIALS
- GROUP SUMS
- VARIABLE PERCENTS
- FACPRINTS
- STS AND PLAN OF INSTRUCTION (POI) REPORTS
- STUDY TABLE OF CONTENTS

FIGURE 4

CODAP/STAT INTERFACE

- CONSOLIDATED OUTPUT
- CODAP BACKGROUND VARIABLES
- CODAP COMPUTED VARIABLES
- IS STATISTICAL VARIABLES

FIGURE 5

SPSS/STAT INTERFACE

- ANOVA
- T-TEST
- BREAKDOWN
- CROSSTABS

FIGURE 6

425

Effects of Absenteeism on the Performance of Air National Guard Personnel

Robert P. Steel and Guy S. Shane
Air Force Institute of Technology

Recent reviews of the absence literature give substance to claims that considerable research on absenteeism has been performed (Mowday, Porter, & Steers, 1982; Muchinsky, 1977). This prodigious body of research has often yielded mixed findings and uncertain conclusions (e.g., Muchinsky, 1977), although some real progress has been made in evaluating traditional common-sense absence frameworks (Scott & Taylor, 1985). For example, Scott and Taylor's meta-analysis has brought some reliable support to the long-standing faith of researchers in job satisfaction - absence relationships.

Traditionally, researchers have regarded absenteeism as a dysfunctional response emitted by an employee attempting to withdraw from an unpleasant work experience. However, this view may be too simplistic; Mowday et al. (1982) note that both **functional** and **dysfunctional** outcomes may be associated with employee absenteeism. However, cases of absence abuse are still considered by most theorists to be primarily dysfunctional behavioral patterns, and their effective control is assumed to pay significant dividends to the organization.

Absence control policies have long been justified on the grounds that absence abuse contributes to labor costs, and it has also been linked conceptually to lower organizational and individual productivity. While evidence of the labor costs associated with absenteeism continues to mount (cf. Mowday et al., 1982), little systematic thought has been given to the purported absence - performance linkage (Mowday et al., 1982). This relationship appears to have been taken for granted.

Common-sense beliefs about absence correlates have increasingly been questioned by researchers and theorists (e.g., Scott & Taylor, 1985). Likewise, the assumed relationship between absenteeism and declining productivity should be subjected to empirical review. Although occasional investigations (Keller, 1984; Sheridan, 1985) have reported incidental correlations between absenteeism and job performance measures, this relationship has not been systematically studied. The present study reports correlations between absenteeism and job performance ratings.

Very little published research has been performed examining sick leave usage rates in the federal sector. This is perhaps unfortunate given the rather liberal sick leave policies currently in existence in the federal government. Leave administration for Department of the Air Force civilian personnel is defined in Air Force Regulation 40-631 (USAF, 1971), and these policies are uniform throughout the entire federal service. The

policy authorizes 13 days of sick leave per year regardless of seniority. There is **no absolute accumulation limit** on the amount of sick leave employees may accrue, and they may "bank" accrued sick leave for other purposes. Accumulated sick leave may be used to retire early without reducing service time annuity payment calculations or to increase the service time used in calculating retirement and survivor annuity payments. In effect, this policy eliminates the "use or loose" incentive contained in many sick leave plans.

Several studies have established links between absence use patterns and organizational policy (Dalton & Perry, 1981; Larson & Fukami, 1985). Policy in force in the federal sector will undoubtedly impact the way sick leave is used, and furthermore, it may impact the way in which absence relates to other organizational outcomes, notably job performance. A second major focus of the study was on federal sector absence - performance relationships.

Method

Sample

Data for the study were provided by 154 full-time employees of an Air National Guard base in the western United States. The typical respondent was male (93%) and between 31 and 40 years old.

Sick Leave

Absenteeism in the present study was operationalized in terms of authorized sick leave usage. The organization provided sick leave data for all employees covering the period January 1, 1983 to December 31, 1983. Absence records for each employee were organized into 26 biweekly pay periods.

Considerable debate has occurred on the selection of an absence metric. Most authors regard frequency indices as psychometrically superior to time lost measures (Chadwick-Jones, Brown, Nicholson, & Sheppard, 1971; Hammer & Landau, 1981; Muchinsky, 1977). A frequency index expresses absence in terms of the number of incidents of absence, regardless of their individual durations. In contrast time lost indices express absence in terms of the amount of time lost (i.e., hours or days lost). Given the impact that choice of index may have on study outcomes, the present investigation employed both a time lost index and a frequency index.

As the present data were collapsed into 26 units of information, this organization imposed some limitation on development of a frequency index. Therefore, frequency represented the number of pay periods in which an absence event occurred, regardless of duration (possible range 0 - 26). Time lost was calculated as the total number of hours lost during the entire year.

Performance Appraisals

Two sources of employee performance appraisals were utilized. Self-appraisals were embedded within a larger survey questionnaire. The self-appraisal method was based on a technique of self-appraisal developed by Steel and Ovalle (1984) called Feedback Based Self-Appraisal. This method is comparable to traditional self-appraisals in most respects except that employees are instructed to base their ratings on feedback they have received from their immediate supervisors. Steel and Ovalle (1984) found Feedback Based Self-Appraisals to be more highly related to supervisory evaluations than were conventional self-ratings. Evaluations were made on five performance dimensions: quantity, quality, efficiency, problem-solving capacity, and adaptability. Seven point agree - disagree rating scales were used.

Supervisory appraisals were solicited from each employee's immediate supervisor. These ratings were made on the same dimensions as the self-ratings. In addition, an overall rating of employee effectiveness was obtained. The supervisory ratings were distributed on 7-point scales ranging from (1) "this employee is far worse than the typical employee" to (7) "this employee is far better than the typical employee."

Procedure

Self- and supervisory appraisals were collected during the summer of 1984. The data were collected in conjunction with a larger research project, and participants were told that their evaluations would be used exclusively for research purposes only. The questionnaire containing the self-evaluations was administered on-site in small group meetings. Respondents were notified that their responses were confidential and anonymous, and they were informed about possible uses to which the data might be put. Social security numbers were requested from participants, and they were used to link appraisal and absence data on a case-by-case basis.

Results

Pearson rs were calculated between the performance appraisal measures and the two absenteeism metrics, time lost and frequency. These correlations are displayed in Table 1. The correlations tended to be small and nonsignificant. Most of them were negative as we anticipated.

No significant correlations were obtained between the self-appraisals and either absence index. These correlations were uniformly quite small.

Correlations between the supervisory ratings of job performance and the two absence metrics produced inconsistent results. The

Table 1

Absence - Performance Correlations

Performance Dimension	Time Lost	Frequency
Self-Appraisal		
Quantity	-.07	-.06
Quality	-.03	-.04
Efficiency	-.05	-.03
Problem-solving	-.01	.00
Adpatability	-.04	-.03
Supervisory Appraisal		
Quantity	-.12	-.06
Quality	.00	-.06
Efficiency	-.15*	-.10
Problem-solving	-.10	-.09
Adaptability	-.19*	-.12
Overall rating	-.10	-.08

Note. Self-appraisal n = 71; supervisory appraisal n = 153.

* $p < .05$

performance - absence frequency correlations were all nonsignificant. However, correlations between the time lost absence measure and two performance dimensions, efficiency and adaptability were statistically significant ($p < .05$). Given the general unreliability and insensitivity usually attributed to performance appraisal measures (e.g., Schmidt, Gast-Rosenberg, & Hunter, 1980) and the unreliability of time lost absence measures (Muchinsky, 1977), the magnitude of these significant correlations probably underestimates the true relationship between absence and performance. These data may reasonably be construed as indicating that excessive sick leave usage is likely to have detrimental effects on the work performance of federal employees.

Discussion

Although at first blush the significant correlations appear to be small and potentially trivial, we believe they probably underestimate true performance - absence relationships. We believe the cards were stacked against finding significant correlations. The purported reliability of both variables is low, and as Hunter, Schmidt, and Jackson (1982) have stridently pointed out, such artifacts as error of measurement contribute substantially to the apparent magnitude of bivariate relationships. Furthermore, we believe the generous sick leave accumulation policy in force in the federal sector removes the "use or loose" incentives associated with most sick leave plans. This policy should have the effect of reducing absence criterion variance because less total sick leave will be taken. Hence, we believe the results give genuine support to the long-standing belief among personnel managers that absence prone employees are also poor performers. However, the correlational nature of the data prohibit any causal inferences. It is impossible to determine whether the performance of sick leave abusers is rated low because they are absent so often (i.e., performance judgments are influenced by general negative impressions) or because their performance genuinely suffers from so much lost work time.

The study yielded a couple of surprises. Many researchers have observed that frequency absence metrics are more reliable than time lost absence scales. The degree of reliability of an absence measure should have a profound effect on the ability of that metric to correlate with other variables. This fact was clearly manifest in Scott and Taylor's (1985) research. However, the present study failed to obtain significant correlations between performance and absence frequency, although significant rs were obtained for the normally inferior time lost index. We suspect that collapsing the data into 26 pay periods reduced the total amount of variance obtainable with the frequency metric, and perhaps this constraint on the potential variance produced the poor results with this absence index.

References

Chadwick-Jones, J. K., Brown, C. A., Nicholson, N., & Sheppard, C. (1971). Absence measures: Their reliability and stability in an industrial setting. Personnel Psychology, 24, 463-470.

Dalton, D. R., & Perry, J. L. (1981). Absenteeism and the collective bargaining agreement: An empirical test. Academy of Management Journal, 24, 425-431.

Hammer, T. H., & Landau, J. (1981). Methodological issues in the use of absence data. Journal of Applied Psychology, 66, 574-581.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Keller, R. T. (1984). The role of performance and absenteeism in the prediction of turnover. Academy of Management Journal, 27, 176-183.

Larson, E. W., & Fukami, C. V. (1985). Employee absenteeism: The role of ease of movement. Academy of Management Journal, 28, 464-471.

Mowday, R. T., Porter, L. W., & Steers, R. M. (1982). Employee-organization linkages: The psychology of commitment, absenteeism, and turnover. New York: Academic Press.

Muchinsky, P. A. (1977). Employee absenteeism: A review of the literature. Journal of Vocational Behavior, 10, 316-340.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

Scott, K. D., & Taylor, G. S. (1985). An examination of conflicting findings on the relationship between job satisfaction and absenteeism: A meta-analysis. Academy of Management Journal, 28, 599-612.

Sheridan, J. E. (1985). A catastrophe model of employee withdrawal leading to low job performance, high absenteeism, and job turnover during the first year of employment. Academy of Management Journal, 28, 88-109.

Steel, R. P., & Ovalle, N. K. 2d. (1984). Self-appraisal based upon supervisory feedback. Personnel Psychology, 37, 667-685.

United States Air Force. (1971). Civilian personnel: Leave administration. AFR 40-630. Washington: HQ USAF.

IN SEARCH OF HONESTY

Allen N. Shub, Ph.D.
Department of Management
Northeastern Illinois University, Chicago, Illinois

In the 4th Century B.C., legend has it that a philosopher named Diogenes wandered about the Athenian countryside in broad daylight with a lighted lantern clutched in his outstretched hand. He was searching for an honest man.

Now, more than 2,300 years later, the search continues as governmental agencies and private industry look for honest employees. In most employment settings today the demand for help is great and often continuous, especially where lower wages have effected maximum turnover and lower reenlistment, where responsibilities are great and temptations are even greater, and where an indifferent society has spawned an atmosphere more prone to dishonesty and violence than its work-ethic-oriented predecessors.

Is it that hard to find an honest and stable employee?

Consider these facts. According to the U.S. Department of Commerce, at least 30% of all business failures each year are due to employee theft and other forms of employee crimes. A study conducted in 1983 concluded that approximately one-third of all employees steal from their employers.

What can be done to combat these losses? To answer this question, it is helpful to understand theft in terms of the profile of a dishonest person and the theft triangle.

Profile of the Dishonest Individual

Research by Dr. William Terris and his associates at London House, Inc., a Chicago-based psychological testing firm, has shown that employee thieves perceive themselves as average persons in a basically dishonest world. In contrast, honest and stable persons perceive themselves as above-average persons in a basically honest world. The employee thief accepts rationalizations of theft behaviors and is tolerant of and less punitive toward theft. If caught, the employee thief responds:

"I did it because...
my supervisor is doing it.
I needed it more than they did.
the company pays me so little for what they get from me.
the company took advantage of me.
it's covered by insurance. No one loses."

In addition, Terris's research has shown that most employee thieves interviewed do not acknowledge their acts of theft as crimes; a criminal is someone who steals more than they do.

The Theft Triangle

Researchers have noted that three elements must be present for an employee to steal: opportunity, need, and attitude.

Opportunity. Most employees have daily opportunities to steal without fear of being caught or punished. In many situations neither apprehension and punishment nor security devices completely deter employee theft. In fact, in some situations, security systems provide the motivated thief with the challenge to be even more creative.

Need. Everyone has needs and wants. Faced with increasing consumer prices (rent, unexpectedly high medical bills, or other expenses), honest employees will either work harder at their jobs in hopes of getting a raise or promotion or they will take on an extra job. The dishonest employee, however, will turn to theft, wherever it is easiest. Often, it is easiest where he or she works.

Attitude. Attitudes determine how needs are satisfied. Honest employees will resist the temptation of theft because of strong moral attitudes.

It would appear that if employers could somehow assess attitudes toward theft, they would be able to screen out individuals prone to theft. In the past, however, employers have used the Diogenes technique--the lighted lantern in broad daylight--and have gotten about the same results. Let's examine some of these traditional hiring techniques in more detail.

Selection Techniques

Good business practice dictates that whatever selection procedure is used in the hiring situation, it should be validated; that is, it should be job-relevant and predictive of future job performance.

Selection Interviews. The interview is widely used in the selection process. It is hard to imagine hiring or recruiting an individual without first having a face-to-face discussion with the applicant.

The typical interviewer relies on personal, subjective values, opinions, and intuitions when making hiring or recruiting decisions. Obviously, there are many pitfalls to this approach. Winning the confidence of an interviewer (and possibly winning the position) can depend upon whether a person is likable, has a neat appearance, and knows how to interview well. This method, however, does not reliably indicate that a person will have the necessary skills and abilities to successfully perform a job. And it certainly does not reliably indicate whether a person is likely to steal on the job. In fact, data on the validity of the interview as a selection device has been poor. Yet, the interview is often the single most used method of selection.

Application Blanks. Nearly all employers require the completion of application blanks. While responses to questions on an application may be

somewhat distorted or exaggerated, falsification is minimized to the extent that an applicant believes that the responses are verifiable. Rarely will an individual admit to leaving a past position because of theft (although one creative individual, later discovered to be a thief, gave as a reason for leaving his last job: "They couldn't afford to keep me").

Resumes. The purpose of a good resume is to sell the applicant to the recruiter. To that end, it is neither in the applicant's best interest, nor reasonable within the format of the typical resume, to include an admission of theft. The trend today is for "formula" resumes, especially with the proliferation of employee outplacement, counseling, and marketing organizations. Certainly, resumes will continue to be used by recruiters, but the resume cannot be expected to be an adequate indicator of future job success or of future employee theft.

Reference and Background Checking. Reference checking generally consists of verification of previous employment, education, personal and business references, and any other information supplied by the applicant that is considered to be job-relevant and deemed important enough by the recruiter to investigate. When the situation warrants, credit and criminal background checks are also made. The name of a reference is provided by an applicant with the expectation that the reference will be favorable. Even when names of references are not specifically provided by an applicant, as in some background investigations, most references tend to be positive. Previous employers are fearful of liability in revealing employee poor performance or dishonesty and are, therefore, very reluctant to supply negative information. Moreover, since some 80% of employee theft goes undetected, previous employers may not even be aware of a former employee's crimes. Reference checks are necessary, but they simply are not enough.

Another pitfall of reference checks and background investigations is that they can be worthless when done superficially by untrained investigators. And many personnel assistants or governmental clerks tend to be just that.

Polygraph Examinations. Diogenes was not the only person in ancient times searching for the honest individual. Both the ancient Chinese and the Arabian Bedouins, pioneers in the development of lie detector tests, believed that the dishonest person's mouth became dry while engaging in deception. To test for dry mouth, the Chinese required that the person chew and spit out rice powder and the Bedouins required that the person lick a hot iron. Dry expectorated powder and burned tongues were considered as evidence of lying. While Diogenes was unsuccessful in his quest for the honest person, presumably some of the Chinese and Bedouins passed this crucial test.

A distinction must be made between two types of dishonesty--lying and stealing. The polygraph was designed to assess untruthfulness. It does so by monitoring the candidate's physiological reactions (pulse rate, blood pressure, respiration, and galvanic skin response) to direct questions. The polygraph is especially useful in facilitating the process of obtaining admissions, often even before the actual polygraph examination has begun.

Besides being costly (often $25 to $50 for preemployment screening), polygraph testing has been banned or restricted as a condition of employment in at least 19 states and has been the subject of continued legislative opposition on many levels. Mechanical truth verification remains highly controversial in legal circles and has generated a great deal of negative publicity. Use of this approach for preemployment screening carries with it the stigma of negative employee relations. In addition, polygraph examinations require skilled examiners, but standards and training vary greatly across the country, thus rendering the results of this screening method highly variable. While possibly effective at screening out those candidates with previous criminal activity, there is no evidence that the polygraph can screen out those who, given the right circumstances and opportunity, are prone to stealing.

Paper-and-Pencil Tests. Until comparatively recently, paper-and-pencil tests for assessing proneness to theft have been virtually nonexistent. Originally developed as a substitute or alternative to the polygraph, paper-and-pencil honesty tests have become entities in their own right.

To validly predict proneness to theft, a good honesty test should assess beliefs about the extent of theft in society, positive attitudes toward theft, ruminations about theft, perceived ease of theft, interthief loyalty, likelihood of detection, knowledge of employee theft, punitiveness, rationalizations about theft, assessments of own honesty, and theft admissions.

In addition, the test should contain a distortion scale, to detect attempts to "fake good" on the test.

All major honesty tests available today attempt, with varying degrees of success, to meet the above requirements in measuring attitudes toward theft. Only two tests, however, appear to be backed up by major, published research: the London House Personnel Selection Inventory and the Reid Report. There are several other tests for which very little technical data are available, including the Stanton Survey, the TA Survey, and the Wilkerson Pre-Employment Audit.

London House Personnel Selection Inventory. The Personnel Selection Inventory (PSI), developed by psychologist William Terris, was first published in 1975. There are several forms of the PSI available to measure attitudes toward and opinions about dishonesty and which include checklists for dollar values of merchandise, property, and money stolen (the admissions section). The only test of its type to do so, the PSI has an independent distortion scale. Other PSI forms available include additional scales for drug abuse, violence, emotional instability, and safety locus of control (accident prevention).

Reid Report. The Reid Report was first published in 1950 by John E. Reid, a polygrapher. The Reid Report consists of two parts: yes/no questions designed to assess the candidate's attitudes about honesty and theft and a set of application-blank-type items, followed by items designed to obtain admissions of theft and other crimes.

Comparison of the Two Major Honesty Tests

There are many ways to compare tests that purport to measure the same traits. Key issues for comparison include:

Ease of Administration. Both the PSI and the Reid Report are rather easily administered. The directions are clear and easy to follow. For both tests, it is necessary that the test be given on company premises under controlled testing situations. The tests should never be given or mailed to applicants for them to take on their own.

Ease of Scoring. The scoring of both tests is controlled closely by the respective publishers. The primary reason for this control is test security; the publishers thus ensure protection of the scoring keys and the norms.

Both publishers provide a mail-in service whereby the client mails in test booklets and receives a report, generally mailed within 24 hours of receipt. Using the mails has obvious disadvantages, of course, for clients needing more immediate results; hence, both publishers offer a telephone method of scoring. In this approach, information from the test is clerically compiled by the test administrator and then phoned in to the publisher's test center; the test data are entered on-line, scored, and reported back to the client within seconds. For telephone scoring, the PSI has the advantage of a simpler preliminary coding than the Reid Report, whose preliminary work is more complicated and requires more manipulation.

On-site computer scoring is another scoring option. The PSI has a software scoring program that can be run on certain personal and mainframe computers, thus providing for immediate results.

Ease of Interpretation. Both tests provide a measure of dishonesty, with assistance from the publishers in determining recommended/not recommended standards for the hiring of candidates. The PSI, in addition, provides an independent distortion scale that indicates the extent to which the candidate is being truthful about the answers given.

Response Format. The Reid Report has mostly yes/no questions, whereas the PSI allows for five to seven or more choices for each test question, thus encouraging more admissions of dishonesty, while reducing applicant stress.

Legality. Both publishers show evidence that their tests are nondiscriminatory and meet various guidelines on employee selection procedures. In addition, the publishers of the PSI and the Reid Report stress in their literature their willingness to assist their clients in any litigation that may be enacted against them.

Research Base. Both the PSI and the Reid Report have published research regarding reliability, validity, fairness, and other technical information. Reid supplies eight publications; London House provides 50 published studies on the PSI.

Review by Buros. Both tests have been reviewed in Buros's Mental Measurements Yearbook. Evaluation of The Reid Report appears in the eighth edition (1978); the review of the PSI is in press for the ninth edition and is currently available in retrieval format from the publishers of Buros. The reader is invited to consult the Buros reviews for additional information.

Concluding Remarks

Paper-and-pencil honesty testing has been somewhat of a controversial area. Psychologists, especially those in the testing profession, have shown a healthy skepticism about whether a paper-and-pencil test can really predict proneness toward employee theft. The major publishers of honesty tests, especially London House and Reid, have countered with a number of well-documented studies that show correlations with polygraph results. For those who are uncomfortable with polygraph examinations as the criterion, London House has also conducted a series of PSI studies with other criteria and methodology--including time series studies (where retail store shrinkage was reduced following the introduction of the PSI for employee screening), predictive studies (where PSI scores were related to future detected thefts), and contrasted groups (where it was found that charity collectors who were more honest as measured by the PSI turned in more money per day then those who scored as less honest). For those who survey the research literature, it is hard not to be convinced of the validity and viability of a paper-and-pencil approach to screening for honesty.

Recently, the field of honesty testing was given even greater legitimacy by a 1984 review by Sackett & Harris in the prestigious Personnel Psychology. Employers have always known that paper-and-pencil ability tests can be highly predictive of future job success. Now the personnel and psychological communities have recognized that paper-and-pencil honesty tests can indeed predict future employee theft.

And, oh yes, remember Diogenes? too bad that honesty tests weren't available in his day. If so, he would not have had the opportunity to immortalize himself as the seeker of honest men. You see, Diogenes and his father were earlier exiled from their native Sinope, reportedly for tampering with the country's currency!

A MODEL OF PSYCHOLOGICAL STRESS AND CONTROL

M.Kastner and W.Neef
University of the Armed Forces,
Faculty for Economics and Organizational Science
Munich, Federal Republic of Germany

Summary

A model for psychological stress which integrates the approaches of McGRATH (1976) and LAZARUS and LAUNIER (1981) is presented. In the context of the above-mentioned model, a further facet of control (emotional/rational) is added to those which have, up to now, been generally accepted (unstable/stable; external/internal, and global/specific); some empirical proof has been found.
Consequences for diagnosis (esp. the construction of questionnaires) and therapeutic measures by means of which optimal stresslevels can be reached are drawn from the linking of stress and rational/emotional control.

1. A simplified version of the model

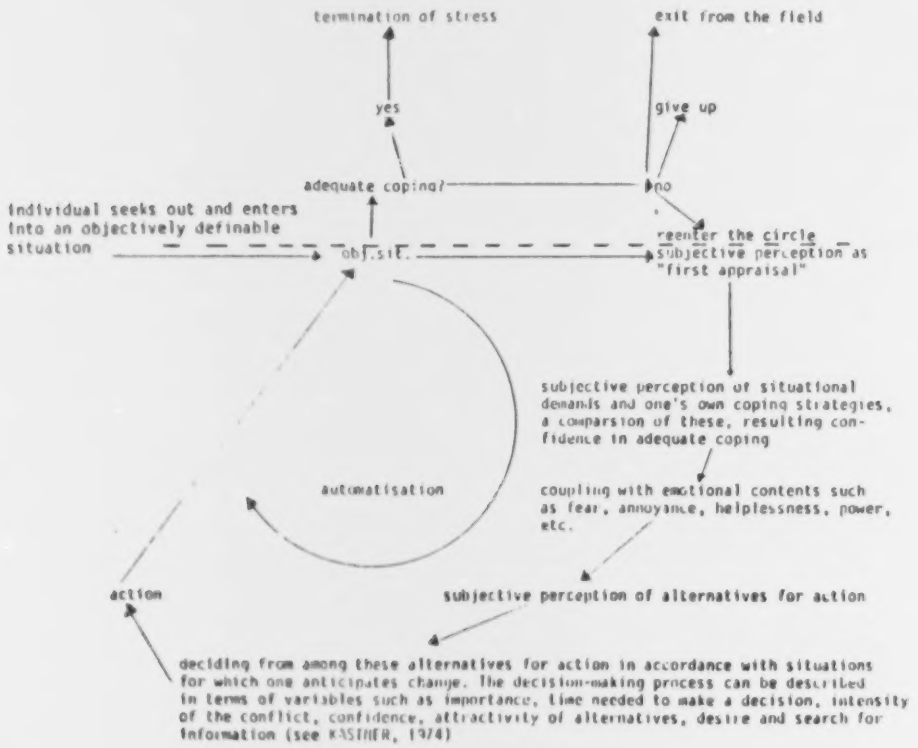
The model consists of four components: one based on a "Handlungs"-theory (i.e. a theory of activity), a transactional one, a part based on facet theory, and the aspect of simultaneous registration of physical and psychological parameters. According to the "Handlungs"-theoretical aspect, the unit of observation must be an action. Such an action can only be described as the dynamic reciprocal action between human and situational components. This introduces the second, the transactional aspect. Reciprocal actions should not be described ideographically in their dazzling uniqueness, but these too should be classified, in order to make diagnosis and measures (including the training of actions) possible. Facet theory is brought into play as a method of classifying situations on the one hand and persons on the other. Objective, physically definable parameters and variables of their psychological subjective representation are needed for these situations. For persons we need objective psychological and physiological (test-) variables as well as variables which subjectively represent a person's own psychological and physiological state. We have herewith arrived at the fourth aspect: how can biological-biochemical and psychological parameters be linked with each other? The goal of diagnosis and of therapy is always to attain an optimal level of

stress, i. e., the oscillation of actions involving stress within an intraindividually varying range of stress.

Later on, the contribution of control to the description of and to predictions concerning the above-mentioned process shall be discussed.

First the model of an individual action:
If the discrepency between subjective situational demands and an individual's own coping potential becomes too great, the individual no longer tries to meet the demand, but instead exits from the field, rationalises, etc. Reflex actions eliminate the right side of the evaluation process. Felxible actions, in the sense of an optimal level of stress, consist in an optimal combination of automatic "small circles" and "large circles" on higher levels of action within the checking and planning process.
Too little stress leads to insufficient experience with these combinations. Too much stress leads to ineffective, uncoordinated combinations, in which normally automated actions rise in hierarchy and therby use valuable cognitive capacity.

Figure 1: Circular process of an action performed under stress (under the interrupted line) and possible ways of terminating this activity (above the interrupted line)


termination of stress
exit from the field
yes
give up
adequate coping?
no
individual seeks out and enters into an objectively definable situation
obj.sit.
reenter the circle
subjective perception as "first appraisal"
subjective perception of situational demands and one's own coping strategies, a comparsion of these, resulting confidence in adequate coping
automatisation
coupling with emotional contents such as fear, annoyance, helplessness, power, etc.
action
subjective perception of alternatives for action
deciding from among these alternatives for action in accordance with situations for which one anticipates change. The decision-making process can be described in terms of variables such as importance, time needed to make a decision, intensity of the conflict, confidence, attractivity of alternatives, desire and search for information (see KASTNER, 1974)

2. Facets of control

How can such an action performed under stress (see figure 1) be influenced by processes of control? Let us first, in the tradition handed down to us by ROTTER (since 1966), take a look at external and internal control. An externally controlled person might possibly focus more on the demands of the situation than on his own potential for coping with it. (The opposite can be said for an internally controlled person.) The comparison between situational demands and a person's own potential for coping with the situation will be strongly dependent on external factors. This may have consequences for the choice of alternatives for action. An externally controlled person will tend to create an alternative for action which he subjectively believes exists. He will most likely "construct" his environment to suit his own taste. The circular model of stress can be similarly played through for the various facets of control which have, up to now, generally been accepted.
Especially when the stress-model was applied to depressive behavior and experience (KASTNER, 1984), it became clear that the "generally accepted" facets of control were not adequate. Questioning test persons again and again made it clear, that depressive persons obviously suffered from a "control split", i. e., the rational elements of control did not correspond to the emotional elements. This led to the splitting up of control into rational and emotional control.
Rationaling control can here be defined as a calculated comparison which is "void of emotion", between the environment and oneself, i. e. the comparison between the stress inherent to a situation and a person's own potential for coping with the situation as well as the appropriate corresponding causal attributions. Emotional control can be defined as the feeling of being in control, of pulling the strings oneself, or of being controlled and having the strings pulled by someone else, i. e., of helplessness.

If we consider only high ad low degrees of both facets, then the following systematic combination results:

		Emotional Control	
		+	-
Rational Control	+	A	B
	-	C	D

Thus four groups of persons are formed, those who are both rationally and emotionally controlled (A), those who are rationally controlled but emotionally uncontrolled (B), those who are rationally uncontrolled but emotionally controlled (C) and finally, persons who are both rationally and emotionally uncontrolled (D).

3. Empirical proof to data

Studies on stress among motorists and persons in a state of depression have shown that

- both facets of control meaningfully describe definable qualities and the existence of the four types refered to above can be proved;
- stress and performance differ for each of the four different "types of control", in accordance with the type and difficulty of the given tasks;
- the best performance for simple speed tasks was given by those who were rationally controlled but emotionally uncontrolled (B);
- the best performance for complex (intelligence) tests was given by those who were both rationally and emotionally controlled.
- The persons in group C accomplished somewhat more in both types of tasks than the persons in group D.
- Persons with a high degree of rational and emotional control experience their actions in stressful traffic situations as being more strongly automated than do uncontrolled persons.
- In addition, rationally and emotionally controlled persons perceive stress situations as being less optically complex. Thus, simplifying the situation and meaningfully integrating it into the context of ones actions seems to be an important function of control.
- With respect to physiological variables, persons who are emotionally controlled but rationally uncontrolled are more excitable than test persons who are both rationally and emotionally uncontrolled (see KASTNER & GUILLOT, 1983).

4. Consequences for future diagnoses and measures

In all of our empirical studies, control as a construct applicable to persons and controllability as an attribute of a situation, proved to be the most predictable variable. The psychological variables we used, taken as a whole, described actions under stress better than did the physiological variables (see KASTNER, 1980; KASTNER & GUILLOT, 1981, 1983). In attempts to link different variables in various manners, it was shown:

- that, first of all, multiplicative linkage of the variables intensity, duration and controllability of the action performed under stress proved more useful than additive linkage;
- that, secondly, control was always the determining variable when it was negatively attributed.
- When control was positively attributed, on the other hand, intensity and, above all, duration, became more important.

Thus, when describing and predicting stress factors, it will be necessary to try to systematically include all facets of control. First of all, controllability must be regarded as peculiar to situations. This contributes to the classification of the situation. Secondly, but of no less importance, control must be regarded as peculiar to persons. In this case, the process in which stress is dealt with will differ, according to whether a person is externally or internally directed, unstable or stable, globally or specifically oriented, emotionally or rationally controlled, or according to how these various possibilities are combined.
The diagnosis can be made using a questionnaire which is constructed on the basis of facet theory, and which systematically varies and combines these facets of control. The therapeutic consequences, on the other hand, should take this differentiation of control into consideration. A person who is rationally controlled but emotionally uncontrolled must be differently treated, taught, trained etc., in consideration of his emotional processes, than a person who is emotionally controlled but rationally uncontrolled.

References

Kastner, M. (1980): Eine Validierung von Beanspruchungsindikatoren im Labor. Cologne· BASt Research Report.

Kastner, M. & Guillot, G. (1981): Beanspruchung von Kraftfahrern im kontrollierten Feld. Cologne: BASt Research Report

Kastner, M. & Guillot, G. (1983): Beanspruchung des Kraftfahrers in Abhängigkeit von Persönlichkeitsmerkmalen. Cologne: BASt Research Report

Kastner, M. (1984): Psychische Beanspruchung und pädagogische Massnahmen im Hinblick auf eine Humanisierung des Arbeitsprozesses. Zeitschrift für erziehungswissenschaftliche Forschung, 18, 131-144

Lazarus, R.S. & Launier, R. (1979): Stress-related Transactions between Person and Environment. New York: McGraw-Hill

McGrath, J.E. (1976): Stress and Behavior in Organizations. In: Dunette, M.H. (ed.): Handbook of Organizational Psychology. New York: Rand McNally

Methodological Considerations in CBI Research

Theodore M. Shlechter and John A. Boldovici
US Army Research Institute, Fort Knox, KY

The purpose of this paper is to examine some methodological considerations in CBI (computer-based instruction) research as applied to military training. We hope that this discussion will help military decision makers in assessing the usefulness and limitations of CBI. We hope also that the issues discussed here will help behavioral scientists and others who provide technical advisory service to the military in assigning priorities to, and evaluating the outcomes of CBI research.

The Armed Services are planning to spend millions in the next few years to expand the use of CBI. The US Army Armor School, for example, is planning to spend over three million dollars in the next three years to develop 350 hours of courseware for various aspects of armor training. The plans include training various levels of people to perform tasks ranging from using volt meters to land navigation and tactical decision-making.

A major source of support for using CBI in military training is the results of research studies which suggest advantages of CBI over conventional instruction. Several reviews of literature on CBI, Kemner-Richardson, Lamos, and West (1984), and Orlansky (1985), for example, have indicated that the use of CBI led to reduced training time and concomitant reductions in price, with little or no decrease in effectiveness. Those reviews also suggest that students favor CBI over conventional instruction. The cited reports have not, however, analyzed possible methodological problems in the reviewed CBI research. Without such analyses one is ill-equipped to make recommendations and decisions about the use of CBI for military training. Our discussion will address a few of the methodological considerations which seem relevant in assessing the relative merits of CBI and conventional instruction as instructional delivery systems.

Methodological Considerations

The methodological considerations in CBI research affect conclusions about costs, attitudes, and effectiveness.

Costs

Orlansky (1985) indicated that the promise of CBI lay in the potential for saving time and money: millions of dollars could be saved by reducing military training time. Orlansky provided the following formula for computing time savings attributable to CBI: Percent Saving = (CBI Time/Conventional Time) x 100. However, for this formula, CBI was being compared to conventional classroom instruction, which was not an equivalent medium to CBI. Avner, Moore, and Smith (1980) have argued that CBI should only be compared to other self-paced individualized instructional media, e.g., programmed texts. Their argument is especially relevant for training time data, because CBI and other forms of individualized instruction allow the students to proceed at their own pace, while conventional instruction does not. Time savings are thus ascribable not to CBI, but to self-pacing, which characterizes nearly all modern instructional innovations. Time savings of approximately 50 percent for both CBI and programmed texts as compared to

conventional instruction have in fact been reported when all three media have been simultaneously tested (Orlansky, 1985). And programmed texts are usually a fraction of CBI's cost for initial implementation.

The costs of alternative media can be examined, not only by comparing training time, but also by estimating life-cycle costs. Kemner-Richardson et al., (1984) noted, for example, that the PLATO system as compared to conventional instruction would eventually lead to $180,000 a year savings. Such long-term cost analyses of CBI systems versus other systems traditionally have involved estimating administrative costs and increases or decreases in the number of faculty which would accompany implementation of new programs. Such analyses, however, have provided only a partial picture of long-term costs. Personnel costs, for instance, also should include expenses associated with training teachers to be "computer-experts." As Shavelson, Winkler, Stasz, Feibel, Robyn, and Shea (1984) observed, using a CBI system in a typical school necessitated a staff development program which trained teachers to be thoroughly knowledgeable in computers. Clearly, such hidden-costs as teacher training programs would reduce the potential savings for CBI.

Cost estimates also should include determining the reliabilities of the compared systems. System reliability is important to measure, because repairing malfunctions will cost money in repairing the problem area and in training time losses. One formula (Frances, Welling, & Levy, 1983) is (Failures Per Hour x Terminals Affected)/(Working Days by Terminals Affected) x 100. The formula, however, only provides a partial index to the educational and financial losses associated with faulty delivery systems. The index does not take into account the potential problems of an implemented system. For the most part, evaluations are conducted for a prototype system which may not relect a system's reliability when implemented. Data should then be collected after the system is fielded and fully implemented.

Evaluators should also collect data on administrative costs after the system has been fully implemented. Orlansky (1985) surprisingly demonstrated that assumptions about administrative costs associated with CBI had rarely been systematically analyzed. For example, very few CBI evaluations have examined the actual expenses (and problems) associated with a CBI system for updating the instructional materials and for hiring the additional personnel needed to operate and maintain the system's equipment. Information obtained when a system is fully implemented is then needed to compare actual costs to estimated costs and correspondingly to determine the validity of certain assumptions about CBI's value for administrative purposes.

Attitudes

Students' attitudes toward CBI may be influenced by instructor views. Clark and Leonard (1985) have suggested that teachers' believe that CBI would help the educational program. One would expect that students would have similar positive views on CBI, because students' attitudes toward an instructional system are influenced by their teachers' attitudes (King, 1975). Determining teachers' attitudes toward CBI is needed to provide insights into students' attitudes. Factors other than the system itself may dictate the students' views.

Instructor obtrusiveness is relevant in evaluating military CBI programs, because instructors usually are involved in the evaluation. Draxl and Aggen (1981) described the need to give briefings to enlist military instructors' interest in, and support of, new instructional systems. Such briefings may have engendered positive attitudes in instructors, while militating against obtaining unbiased reports from students.

Discrepancies between students' responses to questionnaires on the one hand, and more objective data on the other, were found by Shlechter (1985): Soldiers reported that using a light pen was easy, when in fact many of their errors related to problems with using this responding mechanism, such as holding the light pen incorrectly. Additional efforts to compare self-reports with objective performance measures seem warranted.

Shlechter also found that students showed fatigue while completing CBI lessons, but did not report such fatigue on the subjective evaluation questionnaire. Objective measures of fatigue are an important issue in CBI research because the human-factors literature (see McVey, Clauer, & Taylor, 1984) indicates that VDTs (visual display terminals) may be "visual discomfort terminals" with students' not being able to use certain types of terminals for extended periods of time. Evaluators must then measure through objective indexes human factors variables which may affect users' comfort.

Consideration also should be given to examining other factors which may confound attitude reports. Orlansky and String (1981), for example, reported that half the courses which they reviewed lasted one week or less. Such limited duration might not be adequate for ascertaining students' attitudes toward a CBI system. King (1975), and Clark (1985) have noted that most students initially enjoyed working with a CBI program; however, questions remain about a system's ability to sustain students' motivation and enthusiasm once the novelty has worn off.

Effectiveness

Objective comparisons of CBI and other media require that the compared groups be treated identically in every respect except those under investigation. Instructional content, for example, should be the same for the compared groups, as it was in Morrison and Witmer's (1983) comparison of computer-based and print-based job aids. Other media comparisons did not, as noted by Clark (1985), deal with the issue of matching content. For instance, ETS's often cited evaluations of PLATO and TICCIT (Alderman, Appel, & Murphy, 1978) did not include a specific comparison of course materials presented in the compared classroom. Any differences in ETS's evaluations might have been due to presenting different content and not to any features inherent in the CBI system. Evaluators cannot make any claims about the relative effectiveness of CBI as as instructional delivery system unless unconfounded comparisons are made with other media.

Unconfounded comparisons between CBI and other media also require that both educational programs be created with the same degree of effort. Stone (1985) described the extensive effort involved in developing an in-house set of CBI lessons. He showed that developing this courseware involved a team of

References

Chadwick-Jones, J. K., Brown, C. A., Nicholson, N., & Sheppard, C. (1971). Absence measures: Their reliability and stability in an industrial setting. Personnel Psychology, 24, 463-470.

Dalton, D. R., & Perry, J. L. (1981). Absenteeism and the collective bargaining agreement: An empirical test. Academy of Management Journal, 24, 425-431.

Hammer, T. H., & Landau, J. (1981). Methodological issues in the use of absence data. Journal of Applied Psychology, 66, 574-581.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Keller, R. T. (1984). The role of performance and absenteeism in the prediction of turnover. Academy of Management Journal, 27, 176-183.

Larson, E. W., & Fukami, C. V. (1985). Employee absenteeism: The role of ease of movement. Academy of Management Journal, 28, 464-471.

Mowday, R. T., Porter, L. W., & Steers, R. M. (1982). Employee-organization linkages: The psychology of commitment, absenteeism, and turnover. New York: Academic Press.

Muchinsky, P. A. (1977). Employee absenteeism: A review of the literature. Journal of Vocational Behavior, 10, 316-340.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

Scott, K. D., & Taylor, G. S. (1985). An examination of conflicting findings on the relationship between job satisfaction and absenteeism: A meta-analysis. Academy of Management Journal, 28, 599-612.

Sheridan, J. E. (1985). A catastrophe model of employee withdrawal leading to low job performance, high absenteeism, and job turnover during the first year of employment. Academy of Management Journal, 28, 88-109.

Steel, R. P., & Ovalle, N. K. 2d. (1984). Self-appraisal based upon supervisory feedback. Personnel Psychology, 37, 667-685.

United States Air Force. (1971). Civilian personnel: Leave administration. AFR 40-630. Washington: HQ USAF.

IN SEARCH OF HONESTY

Allen N. Shub, Ph.D.
Department of Management
Northeastern Illinois University, Chicago, Illinois

In the 4th Century B.C., legend has it that a philosopher named Diogenes wandered about the Athenian countryside in broad daylight with a lighted lantern clutched in his outstretched hand. He was searching for an honest man.

Now, more than 2,300 years later, the search continues as governmental agencies and private industry look for honest employees. In most employment settings today the demand for help is great and often continuous, especially where lower wages have effected maximum turnover and lower reenlistment, where responsibilities are great and temptations are even greater, and where an indifferent society has spawned an atmosphere more prone to dishonesty and violence than its work-ethic-oriented predecessors.

Is it that hard to find an honest and stable employee?

Consider these facts. According to the U.S. Department of Commerce, at least 30% of all business failures each year are due to employee theft and other forms of employee crimes. A study conducted in 1983 concluded that approximately one-third of all employees steal from their employers.

What can be done to combat these losses? To answer this question, it is helpful to understand theft in terms of the profile of a dishonest person and the theft triangle.

Profile of the Dishonest Individual

Research by Dr. William Terris and his associates at London House, Inc., a Chicago-based psychological testing firm, has shown that employee thieves perceive themselves as average persons in a basically dishonest world. In contrast, honest and stable persons perceive themselves as above-average persons in a basically honest world. The employee thief accepts rationalizations of theft behaviors and is tolerant of and less punitive toward theft. If caught, the employee thief responds:

"I did it because...
my supervisor is doing it.
I needed it more than they did.
the company pays me so little for what they get from me.
the company took advantage of me.
it's covered by insurance. No one loses."

In addition, Terris's research has shown that most employee thieves interviewed do not acknowledge their acts of theft as crimes; a criminal is someone who steals more than they do.

The Theft Triangle

Researchers have noted that three elements must be present for an employee to steal: opportunity, need, and attitude.

Opportunity. Most employees have daily opportunities to steal without fear of being caught or punished. In many situations neither apprehension and punishment nor security devices completely deter employee theft. In fact, in some situations, security systems provide the motivated thief with the challenge to be even more creative.

Need. Everyone has needs and wants. Faced with increasing consumer prices (rent, unexpectedly high medical bills, or other expenses), honest employees will either work harder at their jobs in hopes of getting a raise or promotion or they will take on an extra job. The dishonest employee, however, will turn to theft, wherever it is easiest. Often, it is easiest where he or she works.

Attitude. Attitudes determine how needs are satisfied. Honest employees will resist the temptation of theft because of strong moral attitudes.

It would appear that if employers could somehow assess attitudes toward theft, they would be able to screen out individuals prone to theft. In the past, however, employers have used the Diogenes technique--the lighted lantern in broad daylight--and have gotten about the same results. Let's examine some of these traditional hiring techniques in more detail.

Selection Techniques

Good business practice dictates that whatever selection procedure is used in the hiring situation, it should be validated; that is, it should be job-relevant and predictive of future job performance.

Selection Interviews. The interview is widely used in the selection process. It is hard to imagine hiring or recruiting an individual without first having a face-to-face discussion with the applicant.

The typical interviewer relies on personal, subjective values, opinions, and intuitions when making hiring or recruiting decisions. Obviously, there are many pitfalls to this approach. Winning the confidence of an interviewer (and possibly winning the position) can depend upon whether a person is likable, has a neat appearance, and knows how to interview well. This method, however, does not reliably indicate that a person will have the necessary skills and abilities to successfully perform a job. And it certainly does not reliably indicate whether a person is likely to steal on the job. In fact, data on the validity of the interview as a selection device has been poor. Yet, the interview is often the single most used method of selection.

Application Blanks. Nearly all employers require the completion of application blanks. While responses to questions on an application may be

somewhat distorted or exaggerated, falsification is minimized to the extent that an applicant believes that the responses are verifiable. Rarely will an individual admit to leaving a past position because of theft (although one creative individual, later discovered to be a thief, gave as a reason for leaving his last job: "They couldn't afford to keep me").

Resumes. The purpose of a good resume is to sell the applicant to the recruiter. To that end, it is neither in the applicant's best interest, nor reasonable within the format of the typical resume, to include an admission of theft. The trend today is for "formula" resumes, especially with the proliferation of employee outplacement, counseling, and marketing organizations. Certainly, resumes will continue to be used by recruiters, but the resume cannot be expected to be an adequate indicator of future job success or of future employee theft.

Reference and Background Checking. Reference checking generally consists of verification of previous employment, education, personal and business references, and any other information supplied by the applicant that is considered to be job-relevant and deemed important enough by the recruiter to investigate. When the situation warrants, credit and criminal background checks are also made. The name of a reference is provided by an applicant with the expectation that the reference will be favorable. Even when names of references are not specifically provided by an applicant, as in some background investigations, most references tend to be positive. Previous employers are fearful of liability in revealing employee poor performance or dishonesty and are, therefore, very reluctant to supply negative information. Moreover, since some 80% of employee theft goes undetected, previous employers may not even be aware of a former employee's crimes. Reference checks are necessary, but they simply are not enough.

Another pitfall of reference checks and background investigations is that they can be worthless when done superficially by untrained investigators. And many personnel assistants or governmental clerks tend to be just that.

Polygraph Examinations. Diogenes was not the only person in ancient times searching for the honest individual. Both the ancient Chinese and the Arabian Bedouins, pioneers in the development of lie detector tests, believed that the dishonest person's mouth became dry while engaging in deception. To test for dry mouth, the Chinese required that the person chew and spit out rice powder and the Bedouins required that the person lick a hot iron. Dry expectorated powder and burned tongues were considered as evidence of lying. While Diogenes was unsuccessful in his quest for the honest person, presumably some of the Chinese and Bedouins passed this crucial test.

A distinction must be made between two types of dishonesty--lying and stealing. The polygraph was designed to assess untruthfulness. It does so by monitoring the candidate's physiological reactions (pulse rate, blood pressure, respiration, and galvanic skin response) to direct questions. The polygraph is especially useful in facilitating the process of obtaining admissions, often even before the actual polygraph examination has begun.

Besides being costly (often $25 to $50 for preemployment screening), polygraph testing has been banned or restricted as a condition of employment in at least 19 states and has been the subject of continued legislative opposition on many levels. Mechanical truth verification remains highly controversial in legal circles and has generated a great deal of negative publicity. Use of this approach for preemployment screening carries with it the stigma of negative employee relations. In addition, polygraph examinations require skilled examiners, but standards and training vary greatly across the country, thus rendering the results of this screening method highly variable. While possibly effective at screening out those candidates with previous criminal activity, there is no evidence that the polygraph can screen out those who, given the right circumstances and opportunity, are prone to stealing.

Paper-and-Pencil Tests. Until comparatively recently, paper-and-pencil tests for assessing proneness to theft have been virtually nonexistent. Originally developed as a substitute or alternative to the polygraph, paper-and-pencil honesty tests have become entities in their own right.

To validly predict proneness to theft, a good honesty test should assess beliefs about the extent of theft in society, positive attitudes toward theft, ruminations about theft, perceived ease of theft, interthief loyalty, likelihood of detection, knowledge of employee theft, punitiveness, rationalizations about theft, assessments of own honesty, and theft admissions.

In addition, the test should contain a distortion scale, to detect attempts to "fake good" on the test.

All major honesty tests available today attempt, with varying degrees of success, to meet the above requirements in measuring attitudes toward theft. Only two tests, however, appear to be backed up by major, published research: the London House Personnel Selection Inventory and the Reid Report. There are several other tests for which very little technical data are available, including the Stanton Survey, the TA Survey, and the Wilkerson Pre-Employment Audit.

London House Personnel Selection Inventory. The Personnel Selection Inventory (PSI), developed by psychologist William Terris, was first published in 1975. There are several forms of the PSI available to measure attitudes toward and opinions about dishonesty and which include checklists for dollar values of merchandise, property, and money stolen (the admissions section). The only test of its type to do so, the PSI has an independent distortion scale. Other PSI forms available include additional scales for drug abuse, violence, emotional instability, and safety locus of control (accident prevention).

Reid Report. The Reid Report was first published in 1950 by John E. Reid, a polygrapher. The Reid Report consists of two parts: yes/no questions designed to assess the candidate's attitudes about honesty and theft and a set of application-blank-type items, followed by items designed to obtain admissions of theft and other crimes.

Comparison of the Two Major Honesty Tests

There are many ways to compare tests that purport to measure the same traits. Key issues for comparison include:

Ease of Administration. Both the PSI and the Reid Report are rather easily administered. The directions are clear and easy to follow. For both tests, it is necessary that the test be given on company premises under controlled testing situations. The tests should never be given or mailed to applicants for them to take on their own.

Ease of Scoring. The scoring of both tests is controlled closely by the respective publishers. The primary reason for this control is test security; the publishers thus ensure protection of the scoring keys and the norms.

Both publishers provide a mail-in service whereby the client mails in test booklets and receives a report, generally mailed within 24 hours of receipt. Using the mails has obvious disadvantages, of course, for clients needing more immediate results; hence, both publishers offer a telephone method of scoring. In this approach, information from the test is clerically compiled by the test administrator and then phoned in to the publisher's test center; the test data are entered on-line, scored, and reported back to the client within seconds. For telephone scoring, the PSI has the advantage of a simpler preliminary coding than the Reid Report, whose preliminary work is more complicated and requires more manipulation.

On-site computer scoring is another scoring option. The PSI has a software scoring program that can be run on certain personal and mainframe computers, thus providing for immediate results.

Ease of Interpretation. Both tests provide a measure of dishonesty, with assistance from the publishers in determining recommended/not recommended standards for the hiring of candidates. The PSI, in addition, provides an independent distortion scale that indicates the extent to which the candidate is being truthful about the answers given.

Response Format. The Reid Report has mostly yes/no questions, whereas the PSI allows for five to seven or more choices for each test question, thus encouraging more admissions of dishonesty, while reducing applicant stress.

Legality. Both publishers show evidence that their tests are nondiscriminatory and meet various guidelines on employee selection procedures. In addition, the publishers of the PSI and the Reid Report stress in their literature their willingness to assist their clients in any litigation that may be enacted against them.

Research Base. Both the PSI and the Reid Report have published research regarding reliability, validity, fairness, and other technical information. Reid supplies eight publications; London House provides 50 published studies on the PSI.

Review by Buros. Both tests have been reviewed in Buros's Mental Measurements Yearbook. Evaluation of The Reid Report appears in the eighth edition (1978); the review of the PSI is in press for the ninth edition and is currently available in retrieval format from the publishers of Buros. The reader is invited to consult the Buros reviews for additional information.

Concluding Remarks

Paper-and-pencil honesty testing has been somewhat of a controversial area. Psychologists, especially those in the testing profession, have shown a healthy skepticism about whether a paper-and-pencil test can really predict proneness toward employee theft. The major publishers of honesty tests, especially London House and Reid, have countered with a number of well-documented studies that show correlations with polygraph results. For those who are uncomfortable with polygraph examinations as the criterion, London House has also conducted a series of PSI studies with other criteria and methodology--including time series studies (where retail store shrinkage was reduced following the introduction of the PSI for employee screening), predictive studies (where PSI scores were related to future detected thefts), and contrasted groups (where it was found that charity collectors who were more honest as measured by the PSI turned in more money per day then those who scored as less honest). For those who survey the research literature, it is hard not to be convinced of the validity and viability of a paper-and-pencil approach to screening for honesty.

Recently, the field of honesty testing was given even greater legitimacy by a 1984 review by Sackett & Harris in the prestigious Personnel Psychology. Employers have always known that paper-and-pencil ability tests can be highly predictive of future job success. Now the personnel and psychological communities have recognized that paper-and-pencil honesty tests can indeed predict future employee theft.

And, oh yes, remember Diogenes? too bad that honesty tests weren't available in his day. If so, he would not have had the opportunity to immortalize himself as the seeker of honest men. You see, Diogenes and his father were earlier exiled from their native Sinope, reportedly for tampering with the country's currency!

A MODEL OF PSYCHOLOGICAL STRESS AND CONTROL

M.Kastner and W.Neef
University of the Armed Forces,
Faculty for Economics and Organizational Science
Munich, Federal Republic of Germany

Summary

A model for psychological stress which integrates the approaches of McGRATH (1976) and LAZARUS and LAUNIER (1981) is presented. In the context of the above-mentioned model, a further facet of control (emotional/rational) is added to those which have, up to now, been generally accepted (unstable/stable; external/internal, and global/specific); some empirical proof has been found.
Consequences for diagnosis (esp. the construction of questionnaires) and therapeutic measures by means of which optimal stresslevels can be reached are drawn from the linking of stress and rational/emotional control.

1. A simplified version of the model

The model consists of four components: one based on a "Handlungs"-theory (i.e. a theory of activity), a transactional one, a part based on facet theory, and the aspect of simultaneous registration of physical and psychological parameters. According to the "Handlungs"-theoretical aspect, the unit of observation must be an action. Such an action can only be described as the dynamic reciprocal action between human and situational components. This introduces the second, the transactional aspect. Reciprocal actions should not be described ideographically in their dazzling uniqueness, but these too should be classified, in order to make diagnosis and measures (including the training of actions) possible. Facet theory is brought into play as a method of classifying situations on the one hand and persons on the other. Objective, physically definable parameters and variables of their psychological subjective representation are needed for these situations. For persons we need objective psychological and physiological (test-) variables as well as variables which subjectively represent a person's own psychological and physiological state. We have herewith arrived at the fourth aspect: how can biological-biochemical and psychological parameters be linked with each other? The goal of diagnosis and of therapy is always to attain an optimal level of

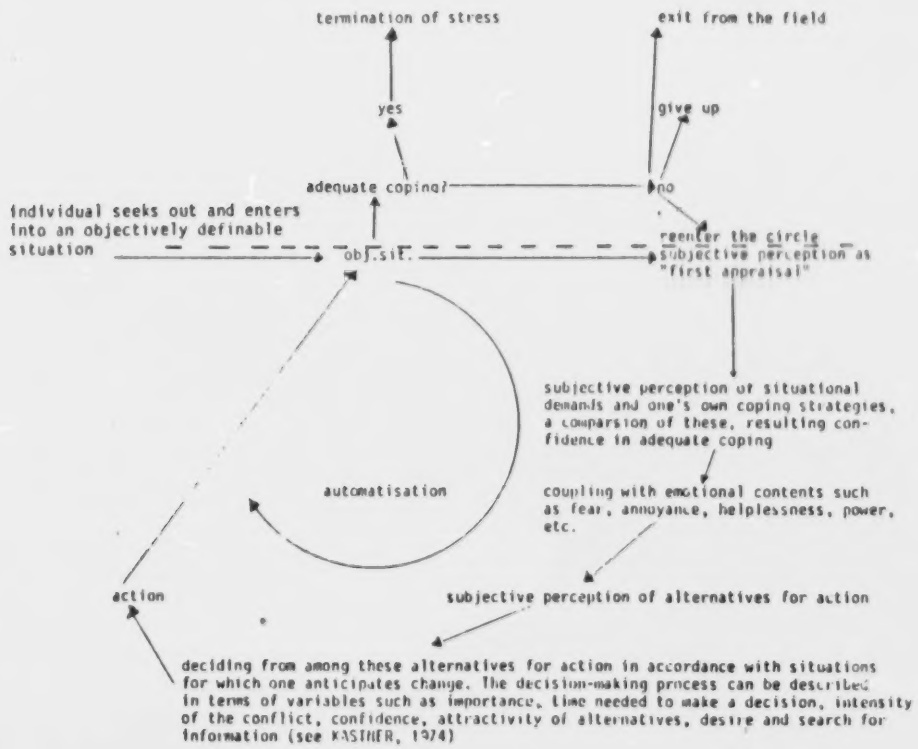
430

stress, i. e., the oscillation of actions involving stress within an intraindividually varying range of stress.

Later on, the contribution of control to the description of and to predictions concerning the above-mentioned process shall be discussed.

First the model of an individual action:
If the discrepency between subjective situational demands and an individual's own coping potential becomes too great, the individual no longer tries to meet the demand, but instead exits from the field, rationalises, etc. Reflex actions eliminate the right side of the evaluation process. Felxible actions, in the sense of an optimal level of stress, consist in an optimal combination of automatic "small circles" and "large circles" on higher levels of action within the checking and planning process.
Too little stress leads to insufficient experience with these combinations. Too much stress leads to ineffective, uncoordinated combinations, in which normally automated actions rise in hierarchy and therby use valuable cognitive capacity.

Figure 1: Circular process of an action performed under stress (under the interrupted line) and possible ways of terminating this activity (above the interrupted line)


termination of stress
exit from the field
yes
give up
adequate coping?
no
individual seeks out and enters into an objectively definable situation
obj.sit.
reenter the circle
subjective perception as "first appraisal"
subjective perception of situational demands and one's own coping strategies, a comparsion of these, resulting confidence in adequate coping
automatisation
coupling with emotional contents such as fear, annoyance, helplessness, power, etc.
action
subjective perception of alternatives for action
deciding from among these alternatives for action in accordance with situations for which one anticipates change. The decision-making process can be described in terms of variables such as importance, time needed to make a decision, intensity of the conflict, confidence, attractivity of alternatives, desire and search for information (see KASTNER, 1974)

2. Facets of control

How can such an action performed under stress (see figure 1) be influenced by processes of control? Let us first, in the tradition handed down to us by ROTTER (since 1966), take a look at external and internal control. An externally controlled person might possibly focus more on the demands of the situation than on his own potential for coping with it. (The opposite can be said for an internally controlled person.) The comparison between situational demands and a person's own potential for coping with the situation will be strongly dependent on external factors. This may have consequences for the choice of alternatives for action. An externally controlled person will tend to create an alternative for action which he subjectively believes exists. He will most likely "construct" his environment to suit his own taste. The circular model of stress can be similarly played through for the various facets of control which have, up to now, generally been accepted.

Especially when the stress-model was applied to depressive behavior and experience (KASTNER, 1984), it became clear that the "generally accepted" facets of control were not adequate. Questioning test persons again and again made it clear, that depressive persons obviously suffered from a "control split", i. e., the rational elements of control did not correspond to the emotional elements. This led to the splitting up of control into rational and emotional control.

Rationaling control can here be defined as a calculated comparison which is "void of emotion", between the environment and oneself, i. e. the comparison between the stress inherent to a situation and a person's own potential for coping with the situation as well as the appropriate corresponding causal attributions. Emotional control can be defined as the feeling of being in control, of pulling the strings oneself, or of being controlled and having the strings pulled by someone else, i. e., of helplessness.

If we consider only high ad low degrees of both facets, then the following systematic combination results:

		Emotional Control	
		+	-
Rational Control	+	A	B
	-	C	D

Thus four groups of persons are formed, those who are both rationally and emotionally controlled (A), those who are rationally controlled but emotionally uncontrolled (B), those who are rationally uncontrolled but emotionally controlled (C) and finally, persons who are both rationally and emotionally uncontrolled (D).

3. Empirical proof to data

Studies on stress among motorists and persons in a state of depression have shown that

- both facets of control meaningfully describe definable qualities and the existence of the four types refered to above can be proved;
- stress and performance differ for each of the four different "types of control", in accordance with the type and difficulty of the given tasks;
- the best performance for simple speed tasks was given by those who were rationally controlled but emotionally uncontrolled (B);
- the best performance for complex (intelligence) tests was given by those who were both rationally and emotionally controlled.
- The persons in group C accomplished somewhat more in both types of tasks than the persons in group D.
- Persons with a high degree of rational and emotional control experience their actions in stressful traffic situations as being more strongly automated than do uncontrolled persons.
- In addition, rationally and emotionally controlled persons perceive stress situations as being less optically complex. Thus, simplifying the situation and meaningfully integrating it into the context of ones actions seems to be an important function of control.
- With respect to physiological variables, persons who are emotionally controlled but rationally uncontrolled are more excitable than test persons who are both rationally and emotionally uncontrolled (see KASTNER & GUILLOT, 1983).

4. Consequences for future diagnoses and measures

In all of our empirical studies, control as a construct applicable to persons and controllability as an attribute of a situation, proved to be the most predictable variable. The psychological variables we used, taken as a whole, described actions under stress better than did the physiological variables (see KASTNER, 1980; KASTNER & GUILLOT, 1981, 1983). In attempts to link different variables in various manners, it was shown:

- that, first of all, multiplicative linkage of the variables intensity, duration and controllability of the action performed under stress proved more useful than additive linkage;
- that, secondly, control was always the determining variable when it was negatively attributed.
- When control was positively attributed, on the other hand, intensity and, above all, duration, became more important.

Thus, when describing and predicting stress factors, it will be necessary to try to systematically include all facets of control. First of all, controllability must be regarded as peculiar to situations. This contributes to the classification of the situation. Secondly, but of no less importance, control must be regarded as peculiar to persons. In this case, the process in which stress is dealt with will differ, according to whether a person is externally or internally directed, unstable or stable, globally or specifically oriented, emotionally or rationally controlled, or according to how these various possibilities are combined.
The diagnosis can be made using a questionnaire which is constructed on the basis of facet theory, and which systematically varies and combines these facets of control. The therapeutic consequences, on the other hand, should take this differentiation of control into consideration. A person who is rationally controlled but emotionally uncontrolled must be differently treated, taught, trained etc., in consideration of his emotional processes, than a person who is emotionally controlled but rationally uncontrolled.

References

Kastner, M. (1980): Eine Validierung von Beanspruchungsindikatoren im Labor. Cologne: BASt Research Report.

Kastner, M. & Guillot, G. (1981): Beanspruchung von Kraftfahrern im kontrollierten Feld. Cologne: BASt Research Report

Kastner, M. & Guillot, G. (1983): Beanspruchung des Kraftfahrers in Abhängigkeit von Persönlichkeitsmerkmalen. Cologne: BASt Research Report

Kastner, M. (1984): Psychische Beanspruchung und pädagogische Massnahmen im Hinblick auf eine Humanisierung des Arbeitsprozesses. Zeitschrift für erziehungswissenschaftliche Forschung, 18, 131-144

Lazarus, R.S. & Launier, R. (1979): Stress-related Transactions between Person and Environment. New York: McGraw-Hill

McGrath, J.E. (1976): Stress and Behavior in Organizations. In: Dunette, M.H. (ed.): Handbook of Organizational Psychology. New York: Rand McNally

Methodological Considerations in CBI Research

Theodore M. Shlechter and John A. Boldovici
US Army Research Institute, Fort Knox, KY

The purpose of this paper is to examine some methodological considerations in CBI (computer-based instruction) research as applied to military training. We hope that this discussion will help military decision makers in assessing the usefulness and limitations of CBI. We hope also that the issues discussed here will help behavioral scientists and others who provide technical advisory service to the military in assigning priorities to, and evaluating the outcomes of CBI research.

The Armed Services are planning to spend millions in the next few years to expand the use of CBI. The US Army Armor School, for example, is planning to spend over three million dollars in the next three years to develop 350 hours of courseware for various aspects of armor training. The plans include training various levels of people to perform tasks ranging from using volt meters to land navigation and tactical decision-making.

A major source of support for using CBI in military training is the results of research studies which suggest advantages of CBI over conventional instruction. Several reviews of literature on CBI, Kemner-Richardson, Lamos, and West (1984), and Orlansky (1985), for example, have indicated that the use of CBI led to reduced training time and concomitant reductions in price, with little or no decrease in effectiveness. Those reviews also suggest that students favor CBI over conventional instruction. The cited reports have not, however, analyzed possible methodological problems in the reviewed CBI research. Without such analyses one is ill-equipped to make recommendations and decisions about the use of CBI for military training. Our discussion will address a few of the methodological considerations which seem relevant in assessing the relative merits of CBI and conventional instruction as instructional delivery systems.

Methodological Considerations

The methodological considerations in CBI research affect conclusions about costs, attitudes, and effectiveness.

Costs

Orlansky (1985) indicated that the promise of CBI lay in the potential for saving time and money: millions of dollars could be saved by reducing military training time. Orlansky provided the following formula for computing time savings attributable to CBI: Percent Saving = (CBI Time/Conventional Time) x 100. However, for this formula, CBI was being compared to conventional classroom instruction, which was not an equivalent medium to CBI. Avner, Moore, and Smith (1980) have argued that CBI should only be compared to other self-paced individualized instructional media, e.g., programmed texts. Their argument is especially relevant for training time data, because CBI and other forms of individualized instruction allow the students to proceed at their own pace, while conventional instruction does not. Time savings are thus ascribable not to CBI, but to self-pacing, which characterizes nearly all modern instructional innovations. Time savings of approximately 50 percent for both CBI and programmed texts as compared to

conventional instruction have in fact been reported when all three media have been simultaneously tested (Orlansky, 1985). And programmed texts are usually a fraction of CBI's cost for initial implementation.

The costs of alternative media can be examined, not only by comparing training time, but also by estimating life-cycle costs. Kemner-Richardson et al., (1984) noted, for example, that the PLATO system as compared to conventional instruction would eventually lead to $180,000 a year savings. Such long-term cost analyses of CBI systems versus other systems traditionally have involved estimating administrative costs and increases or decreases in the number of faculty which would accompany implementation of new programs. Such analyses, however, have provided only a partial picture of long-term costs. Personnel costs, for instance, also should include expenses associated with training teachers to be "computer-experts." As Shavelson, Winkler, Stasz, Feibel, Robyn, and Shea (1984) observed, using a CBI system in a typical school necessitated a staff development program which trained teachers to be thoroughly knowledgeable in computers. Clearly, such hidden-costs as teacher training programs would reduce the potential savings for CBI.

Cost estimates also should include determining the reliabilities of the compared systems. System reliability is important to measure, because repairing malfunctions will cost money in repairing the problem area and in training time losses. One formula (Frances, Welling, & Levy, 1983) is (Failures Per Hour x Terminals Affected)/(Working Days by Terminals Affected) x 100. The formula, however, only provides a partial index to the educational and financial losses associated with faulty delivery systems. The index does not take into account the potential problems of an implemented system. For the most part, evaluations are conducted for a prototype system which may not relect a system's reliability when implemented. Data should then be collected after the system is fielded and fully implemented.

Evaluators should also collect data on administrative costs after the system has been fully implemented. Orlansky (1985) surprisingly demonstrated that assumptions about administrative costs associated with CBI had rarely been systematically analyzed. For example, very few CBI evaluations have examined the actual expenses (and problems) associated with a CBI system for updating the instructional materials and for hiring the additional personnel needed to operate and maintain the system's equipment. Information obtained when a system is fully implemented is then needed to compare actual costs to estimated costs and correspondingly to determine the validity of certain assumptions about CBI's value for administrative purposes.

Attitudes

Students' attitudes toward CBI may be influenced by instructor views. Clark and Leonard (1985) have suggested that teachers' believe that CBI would help the educational program. One would expect that students would have similar positive views on CBI, because students' attitudes toward an instructional system are influenced by their teachers' attitudes (King, 1975). Determining teachers' attitudes toward CBI is needed to provide insights into students' attitudes. Factors other than the system itself may dictate the students' views.

Instructor obtrusiveness is relevant in evaluating military CBI programs, because instructors usually are involved in the evaluation. Draxl and Aggen (1981) described the need to give briefings to enlist military instructors' interest in, and support of, new instructional systems. Such briefings may have engendered positive attitudes in instructors, while militating against obtaining unbiased reports from students.

Discrepancies between students' responses to questionnaires on the one hand, and more objective data on the other, were found by Shlechter (1985): Soldiers reported that using a light pen was easy, when in fact many of their errors related to problems with using this responding mechanism, such as holding the light pen incorrectly. Additional efforts to compare self-reports with objective performance measures seem warranted.

Shlechter also found that students showed fatigue while completing CBI lessons, but did not report such fatigue on the subjective evaluation questionnaire. Objective measures of fatigue are an important issue in CBI research because the human-factors literature (see McVey, Clauer, & Taylor, 1984) indicates that VDTs (visual display terminals) may be "visual discomfort terminals" with students' not being able to use certain types of terminals for extended periods of time. Evaluators must then measure through objective indexes human factors variables which may affect users' comfort.

Consideration also should be given to examining other factors which may confound attitude reports. Orlansky and String (1981), for example, reported that half the courses which they reviewed lasted one week or less. Such limited duration might not be adequate for ascertaining students' attitudes toward a CBI system. King (1975), and Clark (1985) have noted that most students initially enjoyed working with a CBI program; however, questions remain about a system's ability to sustain students' motivation and enthusiasm once the novelty has worn off.

Effectiveness

Objective comparisons of CBI and other media require that the compared groups be treated identically in every respect except those under investigation. Instructional content, for example, should be the same for the compared groups, as it was in Morrison and Witmer's (1983) comparison of computer-based and print-based job aids. Other media comparisons did not, as noted by Clark (1985), deal with the issue of matching content. For instance, ETS's often cited evaluations of PLATO and TICCIT (Alderman, Appel, & Murphy, 1978) did not include a specific comparison of course materials presented in the compared classroom. Any differences in ETS's evaluations might have been due to presenting different content and not to any features inherent in the CBI system. Evaluators cannot make any claims about the relative effectiveness of CBI as as instructional delivery system unless unconfounded comparisons are made with other media.

Unconfounded comparisons between CBI and other media also require that both educational programs be created with the same degree of effort. Stone (1985) described the extensive effort involved in developing an in-house set of CBI lessons. He showed that developing this courseware involved a team of

References

Chadwick-Jones, J. K., Brown, C. A., Nicholson, N., & Sheppard, C. (1971). Absence measures: Their reliability and stability in an industrial setting. Personnel Psychology, 24, 463-470.

Dalton, D. R., & Perry, J. L. (1981). Absenteeism and the collective bargaining agreement: An empirical test. Academy of Management Journal, 24, 425-431.

Hammer, T. H., & Landau, J. (1981). Methodological issues in the use of absence data. Journal of Applied Psychology, 66, 574-581.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Keller, R. T. (1984). The role of performance and absenteeism in the prediction of turnover. Academy of Management Journal, 27, 176-183.

Larson, E. W., & Fukami, C. V. (1985). Employee absenteeism: The role of ease of movement. Academy of Management Journal, 28, 464-471.

Mowday, R. T., Porter, L. W., & Steers, R. M. (1982). Employee-organization linkages: The psychology of commitment, absenteeism, and turnover. New York: Academic Press.

Muchinsky, P. A. (1977). Employee absenteeism: A review of the literature. Journal of Vocational Behavior, 10, 316-340.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

Scott, K. D., & Taylor, G. S. (1985). An examination of conflicting findings on the relationship between job satisfaction and absenteeism: A meta-analysis. Academy of Management Journal, 28, 599-612.

Sheridan, J. E. (1985). A catastrophe model of employee withdrawal leading to low job performance, high absenteeism, and job turnover during the first year of employment. Academy of Management Journal, 28, 88-109.

Steel, R. P., & Ovalle, N. K. 2d. (1984). Self-appraisal based upon supervisory feedback. Personnel Psychology, 37, 667-685.

United States Air Force. (1971). Civilian personnel: Leave administration. AFR 40-630. Washington: HQ USAF.

IN SEARCH OF HONESTY

Allen N. Shub, Ph.D.
Department of Management
Northeastern Illinois University, Chicago, Illinois

In the 4th Century B.C., legend has it that a philosopher named Diogenes wandered about the Athenian countryside in broad daylight with a lighted lantern clutched in his outstretched hand. He was searching for an honest man.

Now, more than 2,300 years later, the search continues as governmental agencies and private industry look for honest employees. In most employment settings today the demand for help is great and often continuous, especially where lower wages have effected maximum turnover and lower reenlistment, where responsibilities are great and temptations are even greater, and where an indifferent society has spawned an atmosphere more prone to dishonesty and violence than its work-ethic-oriented predecessors.

Is it that hard to find an honest and stable employee?

Consider these facts. According to the U.S. Department of Commerce, at least 30% of all business failures each year are due to employee theft and other forms of employee crimes. A study conducted in 1983 concluded that approximately one-third of all employees steal from their employers.

What can be done to combat these losses? To answer this question, it is helpful to understand theft in terms of the profile of a dishonest person and the theft triangle.

Profile of the Dishonest Individual

Research by Dr. William Terris and his associates at London House, Inc., a Chicago-based psychological testing firm, has shown that employee thieves perceive themselves as average persons in a basically dishonest world. In contrast, honest and stable persons perceive themselves as above-average persons in a basically honest world. The employee thief accepts rationalizations of theft behaviors and is tolerant of and less punitive toward theft. If caught, the employee thief responds:

"I did it because...
my supervisor is doing it.
I needed it more than they did.
the company pays me so little for what they get from me.
the company took advantage of me.
it's covered by insurance. No one loses."

In addition, Terris's research has shown that most employee thieves interviewed do not acknowledge their acts of theft as crimes; a criminal is someone who steals more than they do.

The Theft Triangle

Researchers have noted that three elements must be present for an employee to steal: opportunity, need, and attitude.

Opportunity. Most employees have daily opportunities to steal without fear of being caught or punished. In many situations neither apprehension and punishment nor security devices completely deter employee theft. In fact, in some situations, security systems provide the motivated thief with the challenge to be even more creative.

Need. Everyone has needs and wants. Faced with increasing consumer prices (rent, unexpectedly high medical bills, or other expenses), honest employees will either work harder at their jobs in hopes of getting a raise or promotion or they will take on an extra job. The dishonest employee, however, will turn to theft, wherever it is easiest. Often, it is easiest where he or she works.

Attitude. Attitudes determine how needs are satisfied. Honest employees will resist the temptation of theft because of strong moral attitudes.

It would appear that if employers could somehow assess attitudes toward theft, they would be able to screen out individuals prone to theft. In the past, however, employers have used the Diogenes technique--the lighted lantern in broad daylight--and have gotten about the same results. Let's examine some of these traditional hiring techniques in more detail.

Selection Techniques

Good business practice dictates that whatever selection procedure is used in the hiring situation, it should be validated; that is, it should be job-relevant and predictive of future job performance.

Selection Interviews. The interview is widely used in the selection process. It is hard to imagine hiring or recruiting an individual without first having a face-to-face discussion with the applicant.

The typical interviewer relies on personal, subjective values, opinions, and intuitions when making hiring or recruiting decisions. Obviously, there are many pitfalls to this approach. Winning the confidence of an interviewer (and possibly winning the position) can depend upon whether a person is likable, has a neat appearance, and knows how to interview well. This method, however, does not reliably indicate that a person will have the necessary skills and abilities to successfully perform a job. And it certainly does not reliably indicate whether a person is likely to steal on the job. In fact, data on the validity of the interview as a selection device has been poor. Yet, the interview is often the single most used method of selection.

Application Blanks. Nearly all employers require the completion of application blanks. While responses to questions on an application may be

somewhat distorted or exaggerated, falsification is minimized to the extent that an applicant believes that the responses are verifiable. Rarely will an individual admit to leaving a past position because of theft (although one creative individual, later discovered to be a thief, gave as a reason for leaving his last job: "They couldn't afford to keep me").

Resumes. The purpose of a good resume is to sell the applicant to the recruiter. To that end, it is neither in the applicant's best interest, nor reasonable within the format of the typical resume, to include an admission of theft. The trend today is for "formula" resumes, especially with the proliferation of employee outplacement, counseling, and marketing organizations. Certainly, resumes will continue to be used by recruiters, but the resume cannot be expected to be an adequate indicator of future job success or of future employee theft.

Reference and Background Checking. Reference checking generally consists of verification of previous employment, education, personal and business references, and any other information supplied by the applicant that is considered to be job-relevant and deemed important enough by the recruiter to investigate. When the situation warrants, credit and criminal background checks are also made. The name of a reference is provided by an applicant with the expectation that the reference will be favorable. Even when names of references are not specifically provided by an applicant, as in some background investigations, most references tend to be positive. Previous employers are fearful of liability in revealing employee poor performance or dishonesty and are, therefore, very reluctant to supply negative information. Moreover, since some 80% of employee theft goes undetected, previous employers may not even be aware of a former employee's crimes. Reference checks are necessary, but they simply are not enough.

Another pitfall of reference checks and background investigations is that they can be worthless when done superficially by untrained investigators. And many personnel assistants or governmental clerks tend to be just that.

Polygraph Examinations. Diogenes was not the only person in ancient times searching for the honest individual. Both the ancient Chinese and the Arabian Bedouins, pioneers in the development of lie detector tests, believed that the dishonest person's mouth became dry while engaging in deception. To test for dry mouth, the Chinese required that the person chew and spit out rice powder and the Bedouins required that the person lick a hot iron. Dry expectorated powder and burned tongues were considered as evidence of lying. While Diogenes was unsuccessful in his quest for the honest person, presumably some of the Chinese and Bedouins passed this crucial test.

A distinction must be made between two types of dishonesty--lying and stealing. The polygraph was designed to assess untruthfulness. It does so by monitoring the candidate's physiological reactions (pulse rate, blood pressure, respiration, and galvanic skin response) to direct questions. The polygraph is especially useful in facilitating the process of obtaining admissions, often even before the actual polygraph examination has begun.

Besides being costly (often $25 to $50 for preemployment screening), polygraph testing has been banned or restricted as a condition of employment in at least 19 states and has been the subject of continued legislative opposition on many levels. Mechanical truth verification remains highly controversial in legal circles and has generated a great deal of negative publicity. Use of this approach for preemployment screening carries with it the stigma of negative employee relations. In addition, polygraph examinations require skilled examiners, but standards and training vary greatly across the country, thus rendering the results of this screening method highly variable. While possibly effective at screening out those candidates with previous criminal activity, there is no evidence that the polygraph can screen out those who, given the right circumstances and opportunity, are prone to stealing.

Paper-and-Pencil Tests. Until comparatively recently, paper-and-pencil tests for assessing proneness to theft have been virtually nonexistent. Originally developed as a substitute or alternative to the polygraph, paper-and-pencil honesty tests have become entities in their own right.

To validly predict proneness to theft, a good honesty test should assess beliefs about the extent of theft in society, positive attitudes toward theft, ruminations about theft, perceived ease of theft, interthief loyalty, likelihood of detection, knowledge of employee theft, punitiveness, rationalizations about theft, assessments of own honesty, and theft admissions.

In addition, the test should contain a distortion scale, to detect attempts to "fake good" on the test.

All major honesty tests available today attempt, with varying degrees of success, to meet the above requirements in measuring attitudes toward theft. Only two tests, however, appear to be backed up by major, published research: the London House Personnel Selection Inventory and the Reid Report. There are several other tests for which very little technical data are available, including the Stanton Survey, the TA Survey, and the Wilkerson Pre-Employment Audit.

London House Personnel Selection Inventory. The Personnel Selection Inventory (PSI), developed by psychologist William Terris, was first published in 1975. There are several forms of the PSI available to measure attitudes toward and opinions about dishonesty and which include checklists for dollar values of merchandise, property, and money stolen (the admissions section). The only test of its type to do so, the PSI has an independent distortion scale. Other PSI forms available include additional scales for drug abuse, violence, emotional instability, and safety locus of control (accident prevention).

Reid Report. The Reid Report was first published in 1950 by John E. Reid, a polygrapher. The Reid Report consists of two parts: yes/no questions designed to assess the candidate's attitudes about honesty and theft and a set of application-blank-type items, followed by items designed to obtain admissions of theft and other crimes.

Comparison of the Two Major Honesty Tests

There are many ways to compare tests that purport to measure the same traits. Key issues for comparison include:

Ease of Administration. Both the PSI and the Reid Report are rather easily administered. The directions are clear and easy to follow. For both tests, it is necessary that the test be given on company premises under controlled testing situations. The tests should never be given or mailed to applicants for them to take on their own.

Ease of Scoring. The scoring of both tests is controlled closely by the respective publishers. The primary reason for this control is test security; the publishers thus ensure protection of the scoring keys and the norms.

Both publishers provide a mail-in service whereby the client mails in test booklets and receives a report, generally mailed within 24 hours of receipt. Using the mails has obvious disadvantages, of course, for clients needing more immediate results; hence, both publishers offer a telephone method of scoring. In this approach, information from the test is clerically compiled by the test administrator and then phoned in to the publisher's test center; the test data are entered on-line, scored, and reported back to the client within seconds. For telephone scoring, the PSI has the advantage of a simpler preliminary coding than the Reid Report, whose preliminary work is more complicated and requires more manipulation.

On-site computer scoring is another scoring option. The PSI has a software scoring program that can be run on certain personal and mainframe computers, thus providing for immediate results.

Ease of Interpretation. Both tests provide a measure of dishonesty, with assistance from the publishers in determining recommended/not recommended standards for the hiring of candidates. The PSI, in addition, provides an independent distortion scale that indicates the extent to which the candidate is being truthful about the answers given.

Response Format. The Reid Report has mostly yes/no questions, whereas the PSI allows for five to seven or more choices for each test question, thus encouraging more admissions of dishonesty, while reducing applicant stress.

Legality. Both publishers show evidence that their tests are nondiscriminatory and meet various guidelines on employee selection procedures. In addition, the publishers of the PSI and the Reid Report stress in their literature their willingness to assist their clients in any litigation that may be enacted against them.

Research Base. Both the PSI and the Reid Report have published research regarding reliability, validity, fairness, and other technical information. Reid supplies eight publications; London House provides 50 published studies on the PSI.

Review by Buros. Both tests have been reviewed in Buros's Mental Measurements Yearbook. Evaluation of The Reid Report appears in the eighth edition (1978); the review of the PSI is in press for the ninth edition and is currently available in retrieval format from the publishers of Buros. The reader is invited to consult the Buros reviews for additional information.

Concluding Remarks

Paper-and-pencil honesty testing has been somewhat of a controversial area. Psychologists, especially those in the testing profession, have shown a healthy skepticism about whether a paper-and-pencil test can really predict proneness toward employee theft. The major publishers of honesty tests, especially London House and Reid, have countered with a number of well-documented studies that show correlations with polygraph results. For those who are uncomfortable with polygraph examinations as the criterion, London House has also conducted a series of PSI studies with other criteria and methodology--including time series studies (where retail store shrinkage was reduced following the introduction of the PSI for employee screening), predictive studies (where PSI scores were related to future detected thefts), and contrasted groups (where it was found that charity collectors who were more honest as measured by the PSI turned in more money per day then those who scored as less honest). For those who survey the research literature, it is hard not to be convinced of the validity and viability of a paper-and-pencil approach to screening for honesty.

Recently, the field of honesty testing was given even greater legitimacy by a 1984 review by Sackett & Harris in the prestigious Personnel Psychology. Employers have always known that paper-and-pencil ability tests can be highly predictive of future job success. Now the personnel and psychological communities have recognized that paper-and-pencil honesty tests can indeed predict future employee theft.

And, oh yes, remember Diogenes? too bad that honesty tests weren't available in his day. If so, he would not have had the opportunity to immortalize himself as the seeker of honest men. You see, Diogenes and his father were earlier exiled from their native Sinope, reportedly for tampering with the country's currency!

A MODEL OF PSYCHOLOGICAL STRESS AND CONTROL

M.Kastner and W.Neef
University of the Armed Forces,
Faculty for Economics and Organizational Science
Munich, Federal Republic of Germany

Summary

A model for psychological stress which integrates the approaches of McGRATH (1976) and LAZARUS and LAUNIER (1981) is presented. In the context of the above-mentioned model, a further facet of control (emotional/rational) is added to those which have, up to now, been generally accepted (unstable/stable; external/internal, and global/specific); some empirical proof has been found.
Consequences for diagnosis (esp. the construction of questionnaires) and therapeutic measures by means of which optimal stresslevels can be reached are drawn from the linking of stress and rational/emotional control.

1. A simplified version of the model

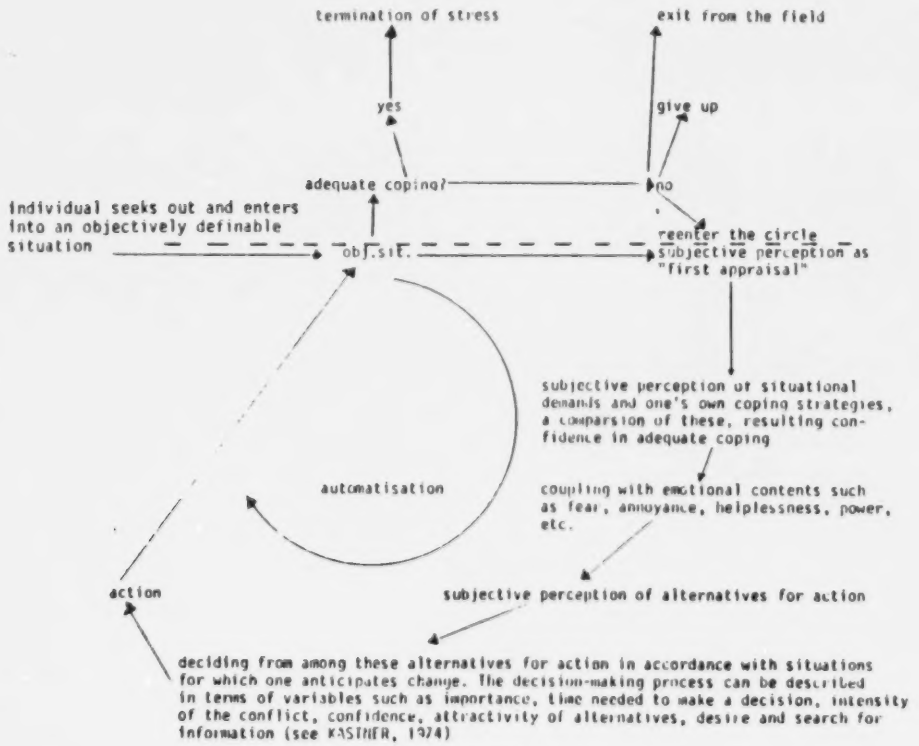
The model consists of four components: one based on a "Handlungs"-theory (i.e. a theory of activity), a transactional one, a part based on facet theory, and the aspect of simultaneous registration of physical and psychological parameters. According to the "Handlungs"-theoretical aspect, the unit of observation must be an action. Such an action can only be described as the dynamic reciprocal action between human and situational components. This introduces the second, the transactional aspect. Reciprocal actions should not be described ideographically in their dazzling uniqueness, but these too should be classified, in order to make diagnosis and measures (including the training of actions) possible. Facet theory is brought into play as a method of classifying situations on the one hand and persons on the other. Objective, physically definable parameters and variables of their psychological subjective representation are needed for these situations. For persons we need objective psychological and physiological (test-) variables as well as variables which subjectively represent a person's own psychological and physiological state. We have herewith arrived at the fourth aspect: how can biological-biochemical and psychological parameters be linked with each other? The goal of diagnosis and of therapy is always to attain an optimal level of

stress, i. e., the oscillation of actions involving stress within an intraindividually varying range of stress.

Later on, the contribution of control to the description of and to predictions concerning the above-mentioned process shall be discussed.

First the model of an individual action:
If the discrepency between subjective situational demands and an individual's own coping potential becomes too great, the individual no longer tries to meet the demand, but instead exits from the field, rationalises, etc. Reflex actions eliminate the right side of the evaluation process. Felxible actions, in the sense of an optimal level of stress, consist in an optimal combination of automatic "small circles" and "large circles" on higher levels of action within the checking and planning process.
Too little stress leads to insufficient experience with these combinations. Too much stress leads to ineffective, uncoordinated combinations, in which normally automated actions rise in hierarchy and therby use valuable cognitive capacity.

Figure 1: Circular process of an action performed under stress (under the interrupted line) and possible ways of terminating this activity (above the interrupted line)


termination of stress
exit from the field
yes
give up
adequate coping?
no
individual seeks out and enters into an objectively definable situation
obj.sit.
reenter the circle
subjective perception as "first appraisal"
subjective perception of situational demands and one's own coping strategies, a comparsion of these, resulting confidence in adequate coping
automatisation
coupling with emotional contents such as fear, annoyance, helplessness, power, etc.
action
subjective perception of alternatives for action
deciding from among these alternatives for action in accordance with situations for which one anticipates change. The decision-making process can be described in terms of variables such as importance, time needed to make a decision, intensity of the conflict, confidence, attractivity of alternatives, desire and search for information (see KASTNER, 1974)

2. Facets of control

How can such an action performed under stress (see figure 1) be influenced by processes of control? Let us first, in the tradition handed down to us by ROTTER (since 1966), take a look at external and internal control. An externally controlled person might possibly focus more on the demands of the situation than on his own potential for coping with it. (The opposite can be said for an internally controlled person.) The comparison between situational demands and a person's own potential for coping with the situation will be strongly dependent on external factors. This may have consequences for the choice of alternatives for action. An externally controlled person will tend to create an alternative for action which he subjectively believes exists. He will most likely "construct" his environment to suit his own taste. The circular model of stress can be similarly played through for the various facets of control which have, up to now, generally been accepted.

Especially when the stress-model was applied to depressive behavior and experience (KASTNER, 1984), it became clear that the "generally accepted" facets of control were not adequate. Questioning test persons again and again made it clear, that depressive persons obviously suffered from a "control split", i. e., the rational elements of control did not correspond to the emotional elements. This led to the splitting up of control into rational and emotional control.

Rationaling control can here be defined as a calculated comparison which is "void of emotion", between the environment and oneself, i. e. the comparison between the stress inherent to a situation and a person's own potential for coping with the situation as well as the appropriate corresponding causal attributions. Emotional control can be defined as the feeling of being in control, of pulling the strings oneself, or of being controlled and having the strings pulled by someone else, i. e., of helplessness.

If we consider only high ad low degrees of both facets, then the following systematic combination results:

		Emotional Control	
		+	-
Rational Control	+	A	B
	-	C	D

Thus four groups of persons are formed, those who are both rationally and emotionally controlled (A), those who are rationally controlled but emotionally uncontrolled (B), those who are rationally uncontrolled but emotionally controlled (C) and finally, persons who are both rationally and emotionally uncontrolled (D).

3. Empirical proof to data

Studies on stress among motorists and persons in a state of depression have shown that

- both facets of control meaningfully describe definable qualities and the existence of the four types refered to above can be proved;
- stress and performance differ for each of the four different "types of control", in accordance with the type and difficulty of the given tasks;
- the best performance for simple speed tasks was given by those who were rationally controlled but emotionally uncontrolled (B);
- the best performance for complex (intelligence) tests was given by those who were both rationally and emotionally controlled.
- The persons in group C accomplished somewhat more in both types of tasks than the persons in group D.
- Persons with a high degree of rational and emotional control experience their actions in stressful traffic situations as being more strongly automated than do uncontrolled persons.
- In addition, rationally and emotionally controlled persons perceive stress situations as being less optically complex. Thus, simplifying the situation and meaningfully integrating it into the context of ones actions seems to be an important function of control.
- With respect to physiological variables, persons who are emotionally controlled but rationally uncontrolled are more excitable than test persons who are both rationally and emotionally uncontrolled (see KASTNER & GUILLOT, 1983).

4. Consequences for future diagnoses and measures

In all of our empirical studies, control as a construct applicable to persons and controllability as an attribute of a situation, proved to be the most predictable variable. The psychological variables we used, taken as a whole, described actions under stress better than did the physiological variables (see KASTNER, 1980; KASTNER & GUILLOT, 1981, 1983). In attempts to link different variables in various manners, it was shown:

- that, first of all, multiplicative linkage of the variables intensity, duration and controllability of the action performed under stress proved more useful than additive linkage;
- that, secondly, control was always the determining variable when it was negatively attributed.
- When control was positively attributed, on the other hand, intensity and, above all, duration, became more important.

Thus, when describing and predicting stress factors, it will be necessary to try to systematically include all facets of control. First of all, controllability must be regarded as peculiar to situations. This contributes to the classification of the situation. Secondly, but of no less importance, control must be regarded as peculiar to persons. In this case, the process in which stress is dealt with will differ, according to whether a person is externally or internally directed, unstable or stable, globally or specifically oriented, emotionally or rationally controlled, or according to how these various possibilities are combined.
The diagnosis can be made using a questionnaire which is constructed on the basis of facet theory, and which systematically varies and combines these facets of control. The therapeutic consequences, on the other hand, should take this differentiation of control into consideration. A person who is rationally controlled but emotionally uncontrolled must be differently treated, taught, trained etc., in consideration of his emotional processes, than a person who is emotionally controlled but rationally uncontrolled.

References

Kastner, M. (1980): Eine Validierung von Beanspruchungsindikatoren im Labor. Cologne· BASt Research Report.

Kastner, M. & Guillot, G. (1981): Beanspruchung von Kraftfahrern im kontrollierten Feld. Cologne: BASt Research Report

Kastner, M. & Guillot, G. (1983): Beanspruchung des Kraftfahrers in Abhängigkeit von Persönlichkeitsmerkmalen. Cologne: BASt Research Report

Kastner, M. (1984): Psychische Beanspruchung und pädagogische Massnahmen im Hinblick auf eine Humanisierung des Arbeitsprozesses. Zeitschrift für erziehungswissenschaftliche Forschung, 18, 131-144

Lazarus, R.S. & Launier, R. (1979): Stress-related Transactions between Person and Environment. New York: McGraw-Hill

McGrath, J.E. (1976): Stress and Behavior in Organizations. In: Dunette, M.H. (ed.): Handbook of Organizational Psychology. New York: Rand McNally

Methodological Considerations in CBI Research

Theodore M. Shlechter and John A. Boldovici
US Army Research Institute, Fort Knox, KY

The purpose of this paper is to examine some methodological considerations in CBI (computer-based instruction) research as applied to military training. We hope that this discussion will help military decision makers in assessing the usefulness and limitations of CBI. We hope also that the issues discussed here will help behavioral scientists and others who provide technical advisory service to the military in assigning priorities to, and evaluating the outcomes of CBI research.

The Armed Services are planning to spend millions in the next few years to expand the use of CBI. The US Army Armor School, for example, is planning to spend over three million dollars in the next three years to develop 350 hours of courseware for various aspects of armor training. The plans include training various levels of people to perform tasks ranging from using volt meters to land navigation and tactical decision-making.

A major source of support for using CBI in military training is the results of research studies which suggest advantages of CBI over conventional instruction. Several reviews of literature on CBI, Kemner-Richardson, Lamos, and West (1984), and Orlansky (1985), for example, have indicated that the use of CBI led to reduced training time and concomitant reductions in price, with little or no decrease in effectiveness. Those reviews also suggest that students favor CBI over conventional instruction. The cited reports have not, however, analyzed possible methodological problems in the reviewed CBI research. Without such analyses one is ill-equipped to make recommendations and decisions about the use of CBI for military training. Our discussion will address a few of the methodological considerations which seem relevant in assessing the relative merits of CBI and conventional instruction as instructional delivery systems.

Methodological Considerations

The methodological considerations in CBI research affect conclusions about costs, attitudes, and effectiveness.

Costs

Orlansky (1985) indicated that the promise of CBI lay in the potential for saving time and money: millions of dollars could be saved by reducing military training time. Orlansky provided the following formula for computing time savings attributable to CBI: Percent Saving = (CBI Time/Conventional Time) x 100. However, for this formula, CBI was being compared to conventional classroom instruction, which was not an equivalent medium to CBI. Avner, Moore, and Smith (1980) have argued that CBI should only be compared to other self-paced individualized instructional media, e.g., programmed texts. Their argument is especially relevant for training time data, because CBI and other forms of individualized instruction allow the students to proceed at their own pace, while conventional instruction does not. Time savings are thus ascribable not to CBI, but to self-pacing, which characterizes nearly all modern instructional innovations. Time savings of approximately 50 percent for both CBI and programmed texts as compared to

conventional instruction have in fact been reported when all three media have been simultaneously tested (Orlansky, 1985). And programmed texts are usually a fraction of CBI's cost for initial implementation.

The costs of alternative media can be examined, not only by comparing training time, but also by estimating life-cycle costs. Kemner-Richardson et al., (1984) noted, for example, that the PLATO system as compared to conventional instruction would eventually lead to $180,000 a year savings. Such long-term cost analyses of CBI systems versus other systems traditionally have involved estimating administrative costs and increases or decreases in the number of faculty which would accompany implementation of new programs. Such analyses, however, have provided only a partial picture of long-term costs. Personnel costs, for instance, also should include expenses associated with training teachers to be "computer-experts." As Shavelson, Winkler, Stasz, Feibel, Robyn, and Shea (1984) observed, using a CBI system in a typical school necessitated a staff development program which trained teachers to be thoroughly knowledgeable in computers. Clearly, such hidden-costs as teacher training programs would reduce the potential savings for CBI.

Cost estimates also should include determining the reliabilities of the compared systems. System reliability is important to measure, because repairing malfunctions will cost money in repairing the problem area and in training time losses. One formula (Frances, Welling, & Levy, 1983) is (Failures Per Hour x Terminals Affected)/(Working Days by Terminals Affected) x 100. The formula, however, only provides a partial index to the educational and financial losses associated with faulty delivery systems. The index does not take into account the potential problems of an implemented system. For the most part, evaluations are conducted for a prototype system which may not relect a system's reliability when implemented. Data should then be collected after the system is fielded and fully implemented.

Evaluators should also collect data on administrative costs after the system has been fully implemented. Orlansky (1985) surprisingly demonstrated that assumptions about administrative costs associated with CBI had rarely been systematically analyzed. For example, very few CBI evaluations have examined the actual expenses (and problems) associated with a CBI system for updating the instructional materials and for hiring the additional personnel needed to operate and maintain the system's equipment. Information obtained when a system is fully implemented is then needed to compare actual costs to estimated costs and correspondingly to determine the validity of certain assumptions about CBI's value for administrative purposes.

Attitudes

Students' attitudes toward CBI may be influenced by instructor views. Clark and Leonard (1985) have suggested that teachers' believe that CBI would help the educational program. One would expect that students would have similar positive views on CBI, because students' attitudes toward an instructional system are influenced by their teachers' attitudes (King, 1975). Determining teachers' attitudes toward CBI is needed to provide insights into students' attitudes. Factors other than the system itself may dictate the students' views.

Instructor obtrusiveness is relevant in evaluating military CBI programs, because instructors usually are involved in the evaluation. Draxl and Aggen (1981) described the need to give briefings to enlist military instructors' interest in, and support of, new instructional systems. Such briefings may have engendered positive attitudes in instructors, while militating against obtaining unbiased reports from students.

Discrepancies between students' responses to questionnaires on the one hand, and more objective data on the other, were found by Shlechter (1985): Soldiers reported that using a light pen was easy, when in fact many of their errors related to problems with using this responding mechanism, such as holding the light pen incorrectly. Additional efforts to compare self-reports with objective performance measures seem warranted.

Shlechter also found that students showed fatigue while completing CBI lessons, but did not report such fatigue on the subjective evaluation questionnaire. Objective measures of fatigue are an important issue in CBI research because the human-factors literature (see McVey, Clauer, & Taylor, 1984) indicates that VDTs (visual display terminals) may be "visual discomfort terminals" with students' not being able to use certain types of terminals for extended periods of time. Evaluators must then measure through objective indexes human factors variables which may affect users' comfort.

Consideration also should be given to examining other factors which may confound attitude reports. Orlansky and String (1981), for example, reported that half the courses which they reviewed lasted one week or less. Such limited duration might not be adequate for ascertaining students' attitudes toward a CBI system. King (1975), and Clark (1985) have noted that most students initially enjoyed working with a CBI program; however, questions remain about a system's ability to sustain students' motivation and enthusiasm once the novelty has worn off.

Effectiveness

Objective comparisons of CBI and other media require that the compared groups be treated identically in every respect except those under investigation. Instructional content, for example, should be the same for the compared groups, as it was in Morrison and Witmer's (1983) comparison of computer-based and print-based job aids. Other media comparisons did not, as noted by Clark (1985), deal with the issue of matching content. For instance, ETS's often cited evaluations of PLATO and TICCIT (Alderman, Appel, & Murphy, 1978) did not include a specific comparison of course materials presented in the compared classroom. Any differences in ETS's evaluations might have been due to presenting different content and not to any features inherent in the CBI system. Evaluators cannot make any claims about the relative effectiveness of CBI as as instructional delivery system unless unconfounded comparisons are made with other media.

Unconfounded comparisons between CBI and other media also require that both educational programs be created with the same degree of effort. Stone (1985) described the extensive effort involved in developing an in-house set of CBI lessons. He showed that developing this courseware involved a team of

Army Captains, civilian educational specialists, and outside consultants working for a year. What typical Army Instructors go to such an effort to develop their daily lesson plans? CBI evaluators must then take into consideration such differences in instructional efforts when discussing their evaluation results. If differences between CBI and conventional instructions only revolve around this issue of instructional effort, then the Army should consider putting forth similar efforts in creating instructional plans to be used by classroom teachers.

CBI evaluators should also determine a system's instructional efficiency by its ability to help the student accomplish different levels of transfer. Clark and Voogel (in press) suggested that CBI was usually geared for immediate transfer and frequently ignored the skills and strategies necessary for long-term retention. They also suggested that CBI courseware was not geared for cognitively oriented tasks, such as problem solving. These trends are disturbing to educators, who place a preminum on long-term mastery of information and the ability to help students develop problem-solving skills. Unfortunately, very few CBI evaluations have examined the long-term and cognitive impact associated with this medium.

The apparent effectiveness of CBI may also be an artifact of unwarranted instructional prompting. Clark and Leonard, after reviewing 42 randomly selected civilian CBI programs, found that teachers usually provided CBI groups with more instructions, e.g. prompts, to complete the tasks than they provided for control students. The extent to which prompting is a problem depends, of course, on the observed amount of prompting relative to the amount intended by the system's designers. Unwarranted prompting should then be another variable measured in CBI research.

Summary and Conclusions

In summary, the following methodological considerations have been discussed regarding CBI research: 1) make unconfounded comparisons between the CBI system and other appropriate educational media; 2) measure hidden life-cycle costs associated with the delivery system; 3) determine the system's reliability; 4) measure actual life-cycle and reliability costs for an implemented system; 5) determine teachers' attitudes toward the CBI system; 6) control for possible "instructor obtrusiveness" effects; 7) substantiate subjective evaluational data with more objective measures; 8) measure human factors variables with objective indexes; 9) control for possible confoundings due to insufficient testing duration; 10) make sure that the compared media have the same content; 11) control for differences in instructional efforts; 12) examine students' long-term and cognitive mastery of the information; and 13) measure unwarranted prompting. Other methodological issues which cannot be described in this paper involve the possible interactions which may exist between student characteristics and the experimental treatment. As argued throughout this paper, clearer answers about CBI's inherent value as a delivery system can be obtained if these considerations are incorporated into the evaluation process. One cannot conclude that CBI is the superior educational medium when confounded comparisons are made with inappropriate media.

These methodological problems reflect the complexities in CBI research. To the extent that CBI research continues to compare CBI to other instructional media, then some time-consuming and expensive research procedures must be employed. For one thing, both cross-sectional and longitudinal data should be collected. A cross-sectional design is needed for initial assessments of students' CBI performance, while longitudinal data are necessary to ascertain the long-term learning associated with the CBI system. Secondly, systematic programs of research are needed in which priorities are assigned to independent variables and variables are systematically manipulated and measured in successive experiments. Such research programs are imperative to analyze systematically as many of the previously cited methodological considerations as possible. Systematic analyses would provide, for example, information about the relationship between estimated and actual CBI life-cycle costs. Programmatic researches needed to provide further insights into whether CBI problems are due to courseware or hardware limitations.

Evaluators should perhaps shift focus from questions of inherent superiority to the identification of the conditions under which CBI and alternative media produce and do not produce desired results. Various media have various strengths which must be first enumerated and then matched with intended instructional settings objectives, and resources. Zemke's (1984) conclusion that CBI may best be used as a supplement to classroom instruction is a case in point.

Military educators should only begin widespread implementation of CBI after clearer answers are provided about this medium's instructional and financial value. History has shown that educational innovations which were implemented without sufficient research and planning were always abandoned for later technological innovations (Montague & Wulfeck, 1984). There is currently some evidence that this abandonment process is beginning to occur for some CBI programs (D Reed, Personal Communications, 6 November 1985). With sizeable financial and personnel investments associated with large-scale CBI implementation, the military can ill afford to continue this historical process. This abandonment process would also be unfortunate because computers--if used properly--could be a valuable instructional tool.

REFERENCES

Alderman, D. L., Appel, L. R., & Murphy, R. T. (1978, April). PLATO and TICCIT: An evaluation of CAI in the community college. Educational Technology, 18, 40-46.

Avner, A., Moore, C., & Smith, S. (1980, May). Active external control: A basis for the superiority of CBI. Journal of Computer-Based Instruction, 6(4), 115-118.

Clark, R. E. (1985). Confounding in educational computing research. Journal of Educational Computing Research, 1(2), 137-147.

Clark, R. E., & Leonard, S. (1985, March). Computer Research Confounding. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Clark, R. E., & Voogel, A. (in press). Transfer of training principles for instructional design. Educational Communication and Technology Journal.

Draxl, M., & Aggen, W. (1981). Computer-based instruction project: A final report. College Park European Division: University of Maryland.

Frances, L. D., & Welling, L. G., & Levy, G. W. (1983). An evaluation of the CAI (Computer-Assisted Instruction) segments of the TOW field-test site and the HAWK CW radar repair course. (TDI-TR-83-3). Fort Monroe, VA: US Army Training and Doctrine Command.

Kemner-Richardson, S., Lamos, J. P., & West, A. S. (1984). The CAI decision handbook. Lowry AFB, CO: US Air Force Human Resources Laboratory.

King, A. T. (1975). Impact of computer-based instructions on attitudes of students and instructors: A review. (AFHRL-TR-75-4). Brooks AFB, TX: US Air Force Systems Command.

McVey, B. W., Clauer, C. K., & Taylor, S. E. A comparison of antiglare filters for positive and negative image displays. (HFL-50). San Jose, CA: IBM Human Factors Center.

Montague, W. E., & Wulfeck II, W. H. (1984, Winter). Computer-based instruction: Will it improve instructional quality? Training Technology Journal, 1(2), 4-19.

Morrison, J. E., & Witmer, B. G. (1983). Comparative evaluation of computer-based and print-based job performance aids. Journal of Computer-Based Instruction, 10(3 & 4), 73-75.

Orlansky, J. (1985, October). The cost-effectiveness of military training. A paper presented at a NATO symposium on Military Training and Cost-Effectiveness of Training. Brussels, Belguim.

Orlansky, J., and String, J. (1981, second quarter). Computer-based instructions for military training. Defense Management Journal, 46-54.

Shavelson, R. J., Winkler, J. P., Stasz, C., Feibel, W., Robyn, A. E., & Shea, S. (1984, March). "Successful teachers" patterns of micro-computer based mathematics and science instructions. Santa Monica, CA: Rand Corporation.

Shlechter, T. M. (1985, November). CBI and students' attitudes. Paper to be presented at the Annual meeting of the Mid-South Educational Research Association, Biloxi, MS.

Stone, C. S. (1985, March). Instruction at the US Army Engineer School: Evolution and management of an in-house coding team. Paper presented at the International Meeting of the Association for the Development of Computer-Based Instructional Systems, Philadelphia.

Zemke, R. (1984, May). Evaluating computer-assisted instruction: The good, bad, and the why. Training, 21, 22-47.

The views expressed in this paper are those of the authors and do not represent the official opinions or policies of the US Army Research Institute or the Department of the Army.

THE EFFECTIVENESS OF COMPUTER GRAPHICS IN VISUAL RECOGNITION TRAINING

BARBARA MCDONALD
AND
BETTY WHITEHILL

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA 92152-6800

Abstract

The effectiveness of computer-based graphics in training a visual recognition skill (radar jamming) was investigated. In recognizing jamming, Naval personnel must differentiate between several types and their visual characteristics. Computer graphics were developed to enhance the features of each type of jamming. The graphics were accompanied by captions which gave the title and a one-sentence description. An experiment was conducted to systematically test the effects of graphics. Subjects were students at the Fleet Combat Training Center, assigned to one of six conditions in a 2x3 design. Subjects received training on animated graphics, with or without captions; still graphics, with or without captions; or captions only; or were assigned to a control group. Subjects were then tested using videotapes of actual jamming. Test results are discussed in terms of the relative merits of animated versus still graphics, and the value of captioning.

Introduction

The use of graphics to enhance the learning process has been of interest to the educational community for years. Research on the use of graphics and text illustrations has shown that they can improve learning by (1) helping students understand what they have read, (2) having pictures substitute for words, and (3) enhancing learner enjoyment. They can also increase retention of information by having more than one way of encoding information (i.e., pictures plus words) and through repetition (Levie & Lentz, 1982).

With the advent of computer based training, graphics have been an even more interesting aspect of instruction since they can be made to move, flash, and present in color. However, the evidence on how critical they are to the success of the instruction is mixed. Since they increase development cost significantly, whether they are necessary and the way in which they are used is of importance (Moore, Nawrocki & Simutis, 1979). It is likely that the importance of graphics depends largely on the task to be learned.

In the present study, graphics were a logical extension of the instruction, the purpose of which was to train visual recognition skills. The graphics were designed to offer the students a series of simplified visual patterns to help organize the learning of important visual characteristics. In the present experiment, two aspects of graphics were explored in addition to the effectiveness of the graphics in general. First, the value of animation was addressed and second, the value of captioning was explored.

Background

The present experiment was designed to investigate the influence of computer-based graphics on recognition of visual patterns. The training and testing of visual recognition skills in Naval

personnel has been of longstanding importance. A critical skill in the Navy is the recognition of electronic countermeasures (ECM), also known as "jamming", on radar systems. In recognizing jamming, several jamming types and their visual characteristics must be differentiated. NPRDC developed a training program to teach the recognition of 11 different types of jamming. This program used computer-based instruction combined with videotapes of jamming. The computer-based training presented the instruction and also presented graphics of the jamming types to enhance the visual characteristics. This training program was extremely successful (see McDonald & Crawford, 1983, 1985).

The computer-based instruction for recognition of jamming was designed to focus on categorical features of the different types of jamming. Each type of jamming has critical visual features which are unique and which are important in defining and identifying the type. In real life, however, jamming can be very difficult to identify because the patterns are complex and can be ambiguous with respect to the critical features. In training, these features were difficult to display in real-life videotape presentations. As a result, the videotapes by themselves were not sufficient to train students to identify each type of jamming. Because of this, graphics were used to highlight the critical features. These graphics were "cartoon-like" renditions, making the features of each jamming type very clear and exaggerated. The literature on concept learning provides evidence that in order to teach students to recognize the prototypical concept, lots of examples must be given and prototypical features must be emphasized. (Gagne & Briggs, 1979). It was the purpose of the present experiment to test the use of the graphics as "feature enhancers" since they were considered extremely helpful in teaching the students by helping the students attend to the features critical to identification of each type. Three questions were of interest in the present experiment: (1) were the graphics effective in teaching the recognition of different types of jamming, (2) were animated graphics more effective than still graphics, and (3) would captions (short verbal descriptions of the jamming types) add to the value of the graphics?

In order to test the effects of graphics without the confounding influence of the accompanying instruction, the experiment was conducted prior to the start of the instruction. For experimental purposes, the pretest to the instruction (which consisted of a videotape with examples of actual jamming), served as the dependent variable. In the experiment, there were 5 graphics conditions used and the students previewed the graphics before they took the pretest and went on to the instruction. The five conditions used were animated graphics, with or without captions; still graphics, with or without captions; and captions alone. These five conditions were compared to a control group that did not receive any pre-instruction. The use of these particular conditions permitted evaluation of whether or not graphics were helpful and also were designed to tease out exactly which aspect of graphics use was the most helpful.

Method

Subjects

Subjects were 114 male enlisted Naval personnel, attending ECCM classes and Advanced Warfare classes at FCTCPAC. Their Navy rates included Operations Specialists (OS), Fire Control Technicians (FT), Electronic Technicians (ET) and Electronic Warfare Technicians (EW). These personnel ranged in rank from E3's to E7's.

A control group was formed from previously collected data in which students attending the same course were given the training as a whole package and did not receive the graphics as a preview condition. 93 students comprised the control group.

Experimental Materials

Two types of graphics were used within the computer-based instruction and were therefore tested in the experiment. In the narrated instruction, the graphics used were still graphics with captions. The captions had a title of the jamming and a short description of what the jamming looked like. In the other part of the training, the graphics were animated with captions. The captions used were the same as used in the still graphics. In order to tease out the contributing factors to the graphics effectiveness, there were five graphics conditions in the experiment. They are as follows:

a) STILL GRAPHICS WITH CAPTIONS: two computer frames of each type of jamming (11 types) with a visual representation (graphic) of the jamming type. Underneath the graphic was the title of the jamming and verbal description of approximately 1 sentence.

b) ANIMATED GRAPHICS WITH CAPTIONS: Two frames of each type of jamming in which the graphic of the jamming type was animated (i.e., the picture presented a moving jamming type on the simulated radar scope). Underneath the graphic was a Title and a verbal description.

c) STILL GRAPHICS WITH TITLE AND NO CAPTIONS: The graphics in this condition were the same used in the STILL GRAPHICS WITH CAPTIONS described above except that underneath the graphic was the title only and no verbal description.

d) ANIMATED GRAPHICS WITH TITLE AND NO CAPTIONS: The graphics in this condition were the same used in the ANIMATED GRAPHICS WITH CAPTIONS except that underneath the graphic was the title only and no verbal description.

e) JAMMING TITLES WITH CAPTIONS AND NO GRAPHICS: In this condition, which served as a "control" to the graphics conditions, there were no graphics presented, only the Jamming titles and verbal descriptions of each type of jamming were presented.

Experimental Procedure

The data for the graphics experiment was collected from May, 1983 to June, 1984. The experiment was designed so that the preview condition would "fit" onto the front end of regular course materials and computer- based instruction. ECM and Advanced Warfare classes were used for this experiment. These classes were each scheduled once a month or every two months. Each class was comprised of 8-10 students. Subjects from each class were randomly assigned to one of the five preview conditions. The students took the preview condition on one day. After a one-day delay students in all conditions took the pretest. After the pretest, they proceeded to complete the computer-based training (which is not of concern in this paper).

Dependent Measures

The dependent measures used in the experiment were the scores obtained on the pretest used in the computer-based instruction. The pretest consisted of videotaped examples of actual jamming. There were 2 videotaped examples of each type of jamming (22 items). In addition, the response time for recognition of each example of jamming was measured in seconds. Finally, for experimental groups, total time spent in the preview condition was measured (in minutes).

Equipment

The microprocessor used for this training was the TERAK, an LSI-11 based dual floppy disk drive system with a 32K work memory capability. It included a keyboard and CRT display for the presentation of black white graphics and text. The computer was used for presentation of instruction and graphics, for presentation of tests and test results, and for data collection. The video equipment used for videotape presentations was a Betamax videotape player and TV monitor.

Results

Table 1 presents means and standard deviations for pretest scores for experimental groups and the control group. An analysis of variance was performed and revealed statistically significant differences between the experimental groups (1-4) and the control groups (5 and 6), $F(1,205) = 11.55, p. < .001$. Further analysis revealed that the animated conditions (groups 2 and 4) had statistically higher scores on the dependent measure than the still graphics conditions (groups 1 and 3), $F(1,92) = 6.17$, $p. < .01$. Finally, according to subsequent tests for differences, the animated with captions group accounted for the significant differences. An analysis of variance between the captions groups versus the no captions groups showed no significant differences, $F(1,92) = 2.82$, $p < .09$.

PREVIEW TEST SCORE MEANS FOR ALL GROUPS			
Condition	N	MEAN	SD
1 Still/Captions	24	48.41	21.69
2 Animated/Captions	20	62.50	14.46
3 Still/No Captions	25	46.24	15.59
4 Animated/No Captions	25	51.16	15.15
5 Captions only	20	40.95	19.58
6 Control	93	43.47	17.89

Table 1. Means and Standard Deviations for Preview Test Scores for all Groups

Response times for recognition of jamming was monitored as well as performance scores. The time in seconds was measured for all instances of recognition from the moment the jamming segment was presented until the computer recorded the student response. Table 2 presents the means and standard deviations for response times for experimental and control groups. Statistical comparisons revealed that groups did not differ significantly in response times, $F(5,201) = 1.00, p. < 0.419$. It is interesting to note that while there were no significant differences, performance was slower for the text only group (5), and the control group. These groups did not have the advantage of previewing the graphic displays of each type of jamming.

PREVIEW TEST RESPONSE SECONDS			
Condition	N	MEAN	SD
1 Still/Captions	24	40.83	14.87
2 Animated/Captions	20	43.90	15.03
3 Still/No Captions	25	40.68	15.23
4 Animated/No Captions	25	41.32	12.74
5 Captions only	20	45.30	10.40
6 Control	93	45.88	15.28

Table 2. Means and Standard Deviations for Response Seconds for all Groups

Table 3 presents the total time (in minutes) spent actually previewing the graphics for the experimental groups. An analysis of variance revealed significant differences between experimental groups, $F(4,109)=18.09, p<.001$. A subsequent test for differences revealed that the animated graphics group with captions (group 2) spent more time viewing the graphics prior to taking the pretest.

PREVIEW TOTAL MEAN MINUTES			
Condition	N	MEAN	SD
1 Still/Captions	24	10.66	1.46
2 Animated/Captions	20	13.10	1.44
3 Still/No Captions	25	9.72	1.20
4 Animated/No Captions	25	11.92	1.28
5 Captions only	20	10.45	1.95

Table 3. Total Mean Minutes Viewing Graphics For All Groups

Conclusions and Discussion

Three questions were of interest in the present experiment. First, the overall effectiveness of graphics in teaching the skill of jamming recognition was explored. Second, the value of animation was tested and finally, the use of captions to accompany the graphics was addressed. The results of the experiment revealed that graphics were helpful and specific information about the use of animation and captions was provided as well.

The graphics groups outperformed the control groups on the pretest, so even though they had not seen actual jamming before, the graphics provided them with enough information to recognize actual instances of the jamming. A more detailed look at the results showed that it was the animated with captions group which accounted for the significant differences. This finding suggests the importance of both animation and captions. Interpretation of this finding must be made with caution, however, because of the fact that there was some similarity between the animated condition and the test condition (they were both moving pictures). Captions appeared to add to the value of the animated condition although there were no main effects for captions. The captions may have provided the animated group with a more efficient way to process and retain the preview information.

Analysis of the time spent previewing the graphics provides another clue to the superior performance of the animated with captions group. This group spent more time in previewing the graphics than did the other graphics groups. Here again, the combination of animation with captions may have given this group more information to process thereby accounting for more time spent. In terms of actual response time in the test itself, there were no significant differences between groups.

The results of this experiment provide some interesting guidelines for the use of graphics. First, graphics appear to be a useful way of giving students advance information about visual characteristics of complex patterns. The findings of this study do suggest that it is important to consider the characteristics of the visual pattern to be learned in developing the graphics. In our case, simplifying the visual pattern

was important. However, including the movement through animation turned out to be critical even though it did add some complexity. Finally, the use of captions cannot be trumpeted based on this experiment but the findings do suggest that they added to the value of the animated graphics.

References

Gagne, R.M. & Briggs, L.L. Principles of Instructional Design. New York: Holt, Rinehart & Winston, 1979.

Levie, W.H. & Lenz, R. Effects of Text Illustration: A Review of the Literature, Educational Communication and Technology Journal, 1982, Vol. 30, 4, 195-232.

McDonald, B.A. & Crawford, A.M. Remote Site Training Using Microprocessors, Journal of Computer-Based Instruction, 1983, Vol.10, 83-87.

McDonald, B.A. & Crawford, A.M. Microprocessor-Based Training: School and Shipboard Evaluations, Navy Personnel Research and Development Center Technical Report 51-85-06, In Press.

Moore, M.V., Nawrocki, L.H. & Simutis, Z.M. The Instructional Effectiveness of Three Levels of Graphics Displays for Computer-Assisted Instruction, U.S. Army Research Institute for the Behavioral and Social Sciences Technical Paper 359, April 1979.

Recent Developments in Job Evaluation Research

Stanley D. Stephenson
School of Business
Southwest Texas State University
San Marcos, Texas 78666

Job evaluation is widely used to establish a specific salary for a specific position. The point system is the most widely used job evaluation procedure. In this system, one or more factors are used to rate each job. Research has typically yielded four factors: Skill, Effort, Responsibility, and Working Conditions. Each of these factors can have one or more subfactors, or scales. For example, under the Skill factor Fraser, Cronshaw, and Alexander (1984) used four scales: Education, Experience, Accuracy, and Complexity. Within each factor (or scale) there are several levels of worth; each level is assigned a certain number of points. Judges (typically experienced employees) then rate each factor, usually based on information contained in a job description, and assign the established number of points. The total number of points across factors equals the job's worth to the organization. All jobs are then ranked according to their point total, and wages are set.

In spite of the wide spread use of job evaluation to determine the salary paid to an employee, recent research has been limited. However, in the late 1970s job evaluation in general came under somewhat of an attack, and consequently some research has been conducted in the last several years.

This research was generated by critical comments concerning the use of job evaluation for determining wages. For instance, one major report (Treiman & Hartmann, 1981) concluded that the reliability of job evaluation methods for determining wage levels has not been established. Treiman (1979) noted that there have been no studies of the validity of job descriptions which in theory form the basis for any job evaluation system. The real impact of these criticisms is that there may now be a "perceived" lack of credibility concerning job evaluation.

More specifically, job evaluation methods came under closer scrutiny due to the comparable worth movement which is rapidly becoming a dominant personnel issue of the 1980s. Comparable worth refers to receiving equal pay for equal job value. Proponents of this movement argue that the historical male-female wage difference is due primarily to a bias inherent in the job evaluation process. Consequently, comparable worth advocates are calling for the use of job evaluation methods that can objectively assess differences in basic job factors, a measurement quality that may not exist in current job evaluation techniques.

The chief criticism of job evaluation is that it is inherently judgmental and therefore possibly biased. Bias can enter into the process at two points: in the writing of the job description and in the evaluation of the job description with respect to the factors/scales selected. In other words, the ranking of jobs and the writing of the job description are basically subjective events.

The sum of the recent questioning of job evaluation techniques has been the publication of several research studies designed to test some of the underlying assumptions, and recent

criticisms of bias in job evaluation. This paper attempts to summarize the findings of these articles.

Rater Reliability

Treiman (1979) has stated that "the reliability of ratings is not particularly encouraging"(p. 40). Doverspike, Carlisi, Barrett, and Alexander (1983) conducted their own review of the literature and found, contrary to Treiman, rather high interrater reliability across studies. They then had 10 raters rate 20 job descriptions, for jobs ranging from clerk to accountant to sales correspondent to supervisor, on a point system job evaluation instrument. The instrument had four factors (skill, effort, responsibility, and working conditions) and a total of 11 scales.

They used a relatively new analytical technique, generalizability analysis. This procedure uses a random effects ANOVA design and permits the analysis of each potential source of error that may affect job ratings. In this study, three facets were analyzed: jobs, scales, and raters. The authors found that the rater factor and its interactions with both scales and jobs produced little variance. With adequate training, sufficient job information, and a properly designed point job evaluation system, the job scores produced by the 10 raters yielded adequate levels of reliability. Moreover, "reliability dropped only slightly when the number of raters, assumed to be from the universe of trained raters, was reduced from 10 to 1"(p. 481). These results certainly do not agree with Treiman's (1979) conclusion; they also question the need for the usual recommendation to have a minimum of 10 raters. Others (e. g., Fraser et al, 1984; Doverspike & Barrett, 1984; Madigan, 1985; and Stephenson, 1985) produced similar results. Also, on a related dimension, neither sex of the job incumbent (Schwab & Grams, 1985) nor sex of the job evaluator (Doverspike et al, 1983; Doverspike & Barrett, 1984) appear to influence final job scores.

In sum, it appears that job evaluation raters contribute very little variance to the job evaluation procedure. If properly trained and adequately informed with respect to the jobs being evaluated, raters can and do reach reliable consensus about the worth of jobs. Moreover, this consensus seems to be obtainable with as little as three raters. Consequently, on the surface raters do not seem to have a biasing impact on job evaluation.

The term, on the surface, was used to preface a finding reported by Madigan (1985). As noted, he found rater reliabilities of at least .85. However, he also reported that the lowest standard error of measurement was $\pm$ 40 points at r = .90. "Hense the 95% confidence interval range of 160 points encompassed four possible classification assignments" (p.145). Doverspike et al (1983) also reported large confidence intervals even with high rater correlations. Error variances this large might be unacceptable for comparable worth evaluation.

Madigan (1985) also reported that in the best of three job evaluation methods he found classification level agreement in only 61 of 120 interrater comparisons. Moreover, he reported classification differences of two or more pay grades in 11% of the cases. Gomez-Mejia, Page, and Tornow (1982) suggested a need to examine evaluation `hit rates' (the percent of cases for whom the estimated and actual grades were the same) as well as the

traditional interrater correlational results. Operationally, these authors considered a position a 'hit' if it was classified by the job evaluation system within ± 1 grade of its assigned grade. They noted that the relationship between correlations and hit rates is uncertain; a correlation of .82 between predicted and actual grade may have a lower hit rate than a correlation of .62. In their best job evaluation method, ± 0 hit rates occurred in 40 percent or fewer of the cases.

Consequently, even though rater reliabilities may be acceptable, this measure may not be capturing the entire picture. Even when correlations are in the .90 range, the percent of actual hits may be unacceptable for actual classification purposes. Madigan (1985) expressed the belief that assessment of potential error variance in evaluation measures must go beyond traditional measures to include an analysis of impact on pay or classification decisions. He also expressed the need to establish acceptable error intervals for determining wages.

Madigan (1985) summarized, "The psychometric adequacy of job worth measures generated by point, guide chart, and PAQ [Position Analysis Questionnaire] evaluation methods is open to serious question. Results of this investigation indicated that previous studies of job evaluation understated potential inconsistency in classification decision making attributable to measurement error." "Consequently none of the three evaluation methods evaluated here exhibited the psychometric qualities desired of a procedure that will serve as the governing criterion in pay classification decisions"(p. 146).

Job Evaluation Techniques

Gomez-Mejia et al (1982) rated management positions using seven job evaluations systems. Similar ratings were reported across methods. "It appears that, given a common job-analysis data base and data-collection tool, various methods used to transform data into a grade prediction can yield essentially comparable results."(p. 806). Madigan (1985) reported that the same job rating decisions were reached across three job evaluation methods. Consequently, both Gomez-Mejia et al (1982) and Madigan (1985) found that job evaluation method apparently does not contribute to job evaluation bias.

Factors/Scales

As noted earlier, Doverspike et al (1983) found little variance due to raters. Scales and jobs and their interactions produced most of the variance reported. Doverspike et al also calculated confidence intervals for scales and jobs and found them to be relatively large. To the authors, this result suggests that a replication of the study or the use of different jobs or scales could produce different results. However, the literature suggests that four factors (skill, effort, responsibility, and working conditions) do seem to be reliable across studies.

These results seem to indicate that the popular job evaluation methods have the capability to distinguish between jobs and that factors/scales do measure different dimensions. Certainly, one would want a job evaluation method both to differentiate between jobs and to have independent scales.

Doverspike and Barrett (1984) conducted a more detailed

analysis of the impact of scales as they relate to possible sex bias in job evaluation results. While they did find high rater reliability, they also reported that reliabilities across scales were not the same. Moreover, (1) the correlations between scale scores and total scores varied across sexes; (2) a factor analysis produced a different set of factors for males versus females; and (3) partial correlations between scale totals and sex group scores produced some scales that favored males and some that favored females. Their results are particularly noteworthy in that, while raters provided reliable results, some scales proved to be biased toward one sex or the other.

Worthy of separate mention is the interpretation provided by Doverspike and Barrett (1984) for scale-sex interaction. They suggested that for male sex-typed jobs complexity of interactions with things was associated with higher skill demands while for female sex-typed jobs greater interaction with people was associated with lower skill demands. "Thus, the worth or meaning of interactions with people and things differed for male and female sex-typed jobs"(p.657).

Literature Summary

Obviously, the job evaluation methods that have seemingly stood the test of time are now coming under closer scrutiny. Rater reliability may be acceptable, but classification hit rates may be unacceptable. Moreover, traditional methods of measuring reliability may overstate rater reliability. Also, the concept of what is an acceptable error rate in classification has not been addressed.

The factors that can be used in job evaluation has reached some consensus; skill, effort, responsibility, and working conditions. The factors themselves do not appear to be biased. However, subfactors (i. e., scales) may not produce similar results for all jobs. Scales that emphasize interacting with things may favor the job evaluation of male jobs while scales that emphasize interacting with people may produce lower scores for female jobs. Obviously, if such scales are not equally represented in a job evaluation method, bias would result.

Implications of Job Evaluation Research for Job Analysis As Conducted in the Armed Services

The natural question arises as to why research on civilian job evaluation methods would be of interest to the Armed Services. Perhaps the most compelling reason is that DoD has always stated that it supports Affirmative Action and Equal Employment Opportunity programs in spirit regardless of whether or not DoD is actually included in Civil Rights legislation. Obviously, comparable worth is an ever expanding part of the civil rights movement. Given that job evaluation and its impact on comparable worth is coming under closer scrutiny in the civilian sector, then the Armed Services should also be interested in these issues. The fact that women are playing an ever expanding role in the Armed Services only adds to this argument.

Another reason is that the Armed Services make extensive use of job analysis data. An interesting trend of the job evaluation research is that it is leading more and more to the issues of the validity and reliability of job descriptions and the underlying data, often job analysis data, that generate these descriptions.

Or, as stated by Gomez-Mejia et al, "... the instrument that is used in gathering job analysis data is a critical element in building a valid and practical job evaluation system"(p. 806).

There are other reasons why the Armed Services should be interested in the results of job evaluation research. First, within the Air Force for example, promotion rates for enlisted personnel are heavily influenced by the results on skill knowledge tests, tests whose questions are directly linked to the task inventories generated in the job analysis process. Second, job descriptions are an important part of the military personnel structure. Since the on-going research in job evaluation is heading towards a more definitive study of job descriptions, the military should also be interested in any validity and reliability research done on job descriptions. Third, we are beginning to witness the inclusion of corresponding civilian job incumbents in job surveys of military specialties. Job evaluation research would be of interest to GS employees. Fourth, from an academic view, the military personnel R & D community should be interested in what the current developments are in the civilian sector both to stay abreast of current developments and also to have the opportunity to add to the investigation of the issues. Related to this fourth reason is the uniqueness of the military environment. Since the military does not have the usual 'wage contamination' found in civilian jobs, it may be able to conduct purer research than is possible in the civilian community.

Suggestions as to the direction military research should take start with the job analysis system (CODAP) because the task inventories used in the CODAP system are the basis for much of the military classification system. The first suggestion is to re-analyze existing CODAP data using recently developed measurement techniques. For example, generalizability analysis could be conducted on job specialties which contain large numbers of female job incumbents. This analysis could take the form of treating individual tasks as job evaluation factor scales and determining the proportion of variance attributable to tasks, scales, factors, or sex group. Partial correlation could also be used to link task inventory responses to group membership. If the task inventories prove to be non-biased in terms of sex preference, then the next step would be to measure the validity of existing job descriptions by having judges independently evaluate jobs based on the current job description and based on the results of the underlying task inventory.

Given that job evaluation research has reported that there is a job by scale variance factor, task factor ratings, such as Job Difficulty and Training Emphasis, should be investigated as well as the number of tasks per job factor. These investigations should study the impact of job factors on job structuring in general as well as the impact, for example, of the job difficulty ratings received by male and female job incumbents.

The basic research goal should be to develop methodologies that can be used both to validate a job analysis task inventory and also to produce and validate subsequent job descriptions. In addition, the Armed Services personnel R & D community should also be able to provide more empirical data on job evaluation classification level confidence intervals and standard errors of

measurement, two stated needs from recent research.

The question is not whether or not current evaluation techniques are working but whether or not they are biased. Current research suggests that there is bias and that this bias begins in either the job description or the data (often job analysis data) that generates the job description. Given the Armed Service's interest in job analysis, it should join in this research. The fact that there has not been much job analysis research done of late by the military is also not an issue; there had not been much job evaluation research done prior to 1979 either. The issue is simply that, just as we need measures of validity of any assessment technique, we need to develop measures of the validity of both (1) the task inventories used in job analysis and (2) the job evaluation methods themselves. The Armed Services can play a major role in this research.

References

Doverspike, D. & Barrett, G. V. (1984). An internal bias analysis of a job evaluation instrument. Journal of Applied Psychology, 69, 648-662.

Doverspike, D., Carlisi, A. M., Barrett, G. V., & Alexander, R. A. (1983). Generalizability analysis of a point-method job evaluation instrument. Journal of Applied Psychology, 68, 476-483.

Fraser, S. L., Cronshaw, S. F., & Alexander, R. A. (1984). Generalizability analysis of a point method job evaluation instrument: A field study. Journal of Applied Psychology, 69, 643-647.

Gomez-Mejia, L. R., Page, R. C., & Tornow, W. W. (1982). A Comparison of the practical utility of traditional, statistical, and hybrid job evaluation approaches. Academy of Management Journal, 25, 790-809.

Madigan, R. M. (1985). Comparable worth judgements: A measurement properties analysis. Journal of Applied Psychology, 70, 137-147.

Schwab, D. P., & Grams, R. (1985). Sex-related errors in job evaluation: A "real world" test. Journal of Applied Psychology, 70, 533-539.

Stephenson, S. D. (1985). Job evaluation for the small business. A paper presented at the 1985 Southwestern Small Business Institute Association Meeting.

Treiman, D. J. (1979). Job evaluation: An analytic review. Washington, DC: National Academy Press.

Treiman, D. J. & Hartman, H. I. (Eds.) (1981). Women, work, and wages: Equal pay for jobs of equal value. Washington, DC: National Academy Press.

PROCEEDINGS

27th Annual Conference of the

MILITARY TESTING ASSOCIATION

Coordinated by the NAVY PERSONNEL RESEARCH & DEVELOPMENT CENTER

SAN DIEGO, CALIFORNIA
21 - 25 OCTOBER 1985

Volume II

530

PROCEEDINGS

27TH ANNUAL CONFERENCE

of the

MILITARY TESTING ASSOCIATION

Coordinated by

the

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

SAN DIEGO, CALIFORNIA

21-25 OCTOBER 1985

VOLUME II

531

CONTENTS

VOLUME II

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Basic Skills Training** Chairperson: Konoske, P.J. Navy Personnel Research and Development Center		
	Pre-Course Mathematics Skills and Success on Advanced Naval Technical Courses	Amyot, CAPT K.A.	462
	Causes of Performance in Air Force Initial Skills Training	Harding, F.D. Mumford, M. Weeks, J.L.	467
PAPER SESSION:	**Methodological and Technological Developments in Survey Research** Chairperson: Morrison, R.F. Navy Personnel Research and Development Center		
	An Examination of the Significance of Non-Participation in Field Studies	Potter, E.H.	473
	Advancing Survey Design through Technology	Doherty, L.M.	480
	Implementation of a Civilian Automated Survey System	Ripkin, F.L. Cecil, S.	486
PAPER SESSION:	**Utilization and Validation of Biodata** Chairperson: Mattson, J.D. Navy Personnel Research and Development Center		
	The Utility of Educational and Biographical Information for Predicting Military Attrition	Riegelhaupt, B.J. Bonczar, T.P.	491

SESSION	TITLE	AUTHOR	PAGE
	Three Variables that May Influence the Validity of Biodata	Walker, C.B.	497
	Use of Personal History Information to Predict Naval Academy Disenrollment	Mattson, J.D. Abrahams, N.M. Hetter, R.D.	503
	Intercorrelations of Biographical Information, Aptitude Test Scores, and Job Performance Ratings for 108 Occupations	Trattner, M.H.	509
PAPER SESSION:	**Training Device Evaluation** Chairperson: Konoske, P.J. Navy Personnel Research and Development Center		
	Analytic Prediction of Training Device Effectiveness	Yates, L.G. Macpherson, D.	515
	Research on Decision Aids for Training Design and Evaluation	Mirabella, A.	520
PAPER SESSION:	**Attitude Surveys** Chairperson: Wilcove, G.L. Navy Personnel Research and Development Center		
	The Army Experience Survey: Methodological Highlights	Celeste, J.F.	526
	USAF Spouse Survey: The Final Chapter	Dansby, MAJ M.R. Ibsen, CPT K.A.	531
	Predictors of Propensity for Continuing Education among Army Chaplains	Pierce, J.E. Goldman, L.A.	537

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Selection Methodogy** Chairperson: Moreno K.E. Navy Personnel Research and Development Center		
	Composite Ratings as a Performance Criterion	Warren, J. Newton, P. Bondaruk, J.	543
	Maximizing Criterion Variance in Validation Research: Is This Always Best?	Angus, CAPT R.J. Ellis, MAJ R.T.	547
SYMPOSIUM:	**Remedial Education in the Navy** Chairperson: Chang, F.R. Navy Personnel Research and Development Center		
	Literacy, Readability and Knowledge	Chang, F.R.	553
	The XFSP: Reading and Mathematics Project	Sticht, T. Armijo, L. Koffman, N. Roberson, K.	559
SYMPOSIUM:	**Elements of a Military Occupational Exploration System** Chairperson: Pass, J.J. Navy Personnel Research and Development Center		
	A Conceptual Framework for Occupational Exploration	Pass, J.J.	565
	Development of the Career Maturity Assessment	Diamond, E.E.	570
	Navy R&D in Support of the Army JOIN System: Leveraging the Government Research Dollar	Baker, H.G. Rafacz, B.A. Sands, W.A.	576

SESSION	TITLE	AUTHOR	PAGE
	Vocational Interests as Predictors of Army Performance	Wing, H. Barge, B.N. Hough, L.M.	582
	Army Career Vocational Guidance as a Recruiting Tool	Hertzbach, A. Knapp, D.J. Johnson, R.M.	587
SYMPOSIUM:	**Determining Ability Requirements** Chairperson: Abrabian, J.M. U.S. Army Research Institute		
	Developing New Attribute Requirements Scales for Military Jobs	Smith, E.P.	593
	Methodological Problems in Identifying Ability Requirements Related to Soldier Performance	Miller, C.R.	599
	Comparison of Weapon Systems Using Ability Requirements Scales	Arabian, J.M.	603
	Computerized Approaches for Estimating Ability Requirements	Rossmeissl, P.G.	609
PAPER SESSION:	**Team Performance Measurement** Chairperson: Flaningam, M.R. Navy Personnel Research and Development Center		
	Chaparral Crew Performance in the Realistic Air Defense Engagement System	Sarli, G.G. Johnson, D.M. Lockhart, J.M.	615
	Command and Control Teams: Techniques for Assessing Team Performance	Cooper, M. Shiflett, S. Korotkin, A.L.	621

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Analysis of the ASVAB** Chairperson: Larson, G.E. Navy Personnel Research and Development Center		
	Item Factor Analysis of ASVAB 14	McCormick, C.	627
	Alternate Forms Reliability of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10	Stern, B.M. White, L.A. Wing, H. Sachs, S.A.	632
PAPER SESSION:	**Personality Assessment** Chairperson: Bjerke, CDR D.G. Navy Personnel Research and Development Center		
	Psychometric Properties of the Safety Locus of Control Scale	Jones, J.W. Wuebker, L.J.	637
	Health Locus of Control Beliefs among Infantrymen	Grieger, L.W.	644
	Classifying Military Offenders: Application of the Megargee MMPI Typology	Paris, M.L. Brown, G.E.	650
PAPER SESSION:	**Team Training** Chairperson: Flaningam, M.R. Navy Personnel Research and Development Center		
	Automated Scheduling of Army Unit Training	Goehring, D.J. Hart, R.J.	655
	Supporting the Changing Role of Army Collective Training Developers	Meliza, L.L.	661
	Application of Model Aircrew Training System (MATS) to B-52 Combat Crew Training	Bills, MAJ C.G. Nullmeyer, R.T.	666

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Analysis and Interpretation of the ASVAB** Chairperson: Moreno, K.E. Navy Personnel Research and Development Center		
	Sensitivities of Speeded Subtests	Wegner, T.G. Ree, M.J.	672
	Initial Operational Test and Evaluation of Armed Service Vocational Aptitude Battery (ASVAB) Forms 11, 12, and 13: Data Quality Analysis	Welsh, J.R. Wegner, T.G.	676
PAPER SESSION:	**Integrated Selection Systems** Chairperson: Bruni, J.R. Navy Personnel Research and Development Center		
	Development of an Integrated Pilot Selection System	Kantor, J.E.	680
	Selection of Skilled Maintenance Employees in the U.S. Postal Service	Mueller, H.A.	686
	The Effects of the Flight Screening Program on Attrition in Undergraduate Pilot Training	Quebe, J.C.	695
PAPER SESSION:	**Issues in Test Item Design** Chairperson: Flaningam, M.R. Navy Personnel Research and Development Center		
	What Performance Does a Performance Test Test?	Ansbro, T.M.	701
	How Qualitatively Informative are Test Items?: A Dense Item Analysis	Bart, W.M.	707
	Test-item Readability: How the Variables Work	Duncan, CPT R.E.	713

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Entry Standards Based on the ASVAB** Chairperson: Vicino, F.L. Navy Personnel Research and Development Center		
	A Comparison of Service Job Standards for Four Military Specialties	Waters, B.K.	719
	Exploring a Statistically Viable Assignment Basis Using ASVAB	Schwartz, M.M.	725
	The Validity of ASVAB for Predicting Training and SQT Performance	Rossemeissl, P.G. McLaughlin, D.H. Wise, L.L. Brandt, D.A.	730
PAPER SESSION:	**Military Attrition Research** Chairperson: Cory, C.H. Navy Personnel Research and Development Center		
	Gender, Ethnic Group, Aptitude and Personality Determinants of U.S. Coast Guard Attrition	Frey, R.L.	736
	Study of Wastage from the Territorial Army (UK Army Reserve)	Blyth, D.M.	741
	Diminishing Returns from an Assessment Centre	Hardy, G.R.	747

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Officer Training and Testing** Chairperson: Greebler, C.S. Navy Personnel Research and Development Center		
	Increasing Basic Skill Levels of Moderate Risk Officer Candidates Prior to Commissioning in Historically Black Colleges	Sprenger, W.D.	753
	Direct Mesurement of ROTC Cadets' Writing Skills	Hanlon, J.P.	759
	Procedures for Refining Written Measurements at USAF OTS	Sako, S. Slaughter, LTC W.J.	763
PAPER SESSION:	**Job Performance Measurement I** Chairperson: Kroeker, L.P. Navy Personnel Research and Development Center		
	Utility Estimation in Five Enlisted Occupations	Eaton, N.K. Wing, H. Lau, A.	769
	Relation of Mental and Education Levels to Navy Enlisted Performance	Cory, C.H.	775
	Selecting Critical Tasks for a Radioman Hands-on Performance Test	Baker, H.G. Laabs, G.J.	780
	Cognitive Predictors of M1 Tank Gunner Performance	Black, B.A.	786

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Issues in Instruction I** Chairperson: Windisch, D.J. Navy Personnel Research and Development Center		
	A Retrospective Analysis of Instructional Technology Innovation Using General System Theory	Ekstrom, A.G.	792
	A Systems Approach to Evaluating High Technology Training	Atwood, N.K.	798
	An Analysis of Attitudes toward Instruction among Vocational Education Instructors	Usova, G.M.	804
PAPER SESSION:	**Assessment of English-as-a-Second Language Personnel** Chairperson: Hetter, R.D. Navy Personnel Research and Development Center		
	Relationship of an Experimental Hispanic Enlistment Screening Test to AFQT	Mathews, J.J. French, C.M.	809
	Attrition and Performance Ratings of ESL Soldiers in BT	Rosenbaum, H.	815
PAPER SESSION:	**Job Performance Measurement II** Chairperson: Kroeker, L.P. Navy Personnel Research and Development Center		
	Standard Setting Methods for Skills Qualification Tests (SQTs)	Pettie, A.L.	821
	Assessing Tank Commander and Gunner Proficiency on U-COFT	Graham, S.E. Boldovici, J.A.	826
	Criterion Referenced Testing for the U.S. Navy's Nuclear Submarine Fleet	Cantor, J.A. Walker, L.	832

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	**Issues in Instruction II** Chairperson: Windisch, D.J. Navy Personnel Research and Development Center		
	A Study of Suggestopedia at the Defense Language Institute Foreign Language Center (DLIFLC)	Bush, B.J.	838
	Heat, Chemical Protective Clothing and Sustained Cognitive Performance	Fine, B.J. Kobrick, J.L.	843
SYMPOSIUM:	**Exceptional Recruits: A Look at High and Low Aptitude Personnel** Chairperson: Lancaster, A.R. Office of the Assistant Secretary of Defense (FM&P)		
	Overview	Lancaster, A.R.	849
	Using Military High- and Low- Aptitude Recruits Wisely	Sellman, W.S.	852
	When Low-Aptitude Recruits Succeed	Means, B. Nigam, A. Heisey, J.G.	855
	Enlistment and Utilization of High-Aptitude Recruits	Laurence, J.H. Schneider, E.F.	861
SYMPOSIUM:	**Predicting a Broad Variety of Criteria: Elaborating the Predictor Space** Chairperson: Wing, H. U.S. Army Research Institute		
	Mapping Predictors to Criterion Space: Overview	Peterson, N.G.	867
	Using Microcomputers for Assessment: Practical Problems and Solutions	Rosse, R.L. Peterson, N.G.	873

SESSION	TITLE	AUTHOR	PAGE
	Computerized Assessment of Perpetual and Psychomotor Abilities	McHenry, J.J. Toquam, J.L.	879
	Adding to the ASVAB: Cognitive Paper-and-Pencil Measures	Toquam, J.L. Dunnette, M.D. Corpe, V.A. Houston, J.	885
	Measuring Personnel Attributes: Temperament, Biodata and Interest	Hough, L.M. McGue, M.K. Kamp, J.D. Houston, J.S. Barge, B.N.	891
SYMPOSIUM:	**The Training and Selection of Army Managers: Quantitative/Qualitative Approaches** Chairperson: Fisher, G.P. Human Resources Research Organization		
	Overview	Fisher, G.P. Lilienthal, R.	897
	Selecting and Training Logistics Professionals and Managers: Qualitative/ Quantitative Approaches	Fisher, G. Lilienthal, R. Hough, L.	900
	Managerial Competencies Assessment For Army Civilians	King, G. C.	906
SYMPOSIUM:	**Army Research Institute R&D Program on the National Training Center** Chairperson: Banks, J.H. U.S. Army Research Institute		
	Types and Quality of National Training Center Data	Whitmarsh, P.J.	912
	An Overview of ARI's Research Program on the National Training Center	Banks, J.H.	916

SESSION	TITLE	AUTHOR	PAGE
	Leader Performance Criteria at the National Training Center (NTC)	Pence, E.C.	922
	Battalion Performance on the Live Fire Range at the National Training Center (NTC)	Forsythe, T.K. Doherty, W.J.	928
	Analysis of NTC Force-on-Force Performance	Nichols, J.J. Doherty, W.J.	931
	Use of Instrumentation to Improve Combat Readiness	Shackelford, W.L.	937
SYMPOSIUM:	**Two Systems for Computer Analysis of Student Feedback** Chairperson: Ekwall, R.W. U.S. Army Command and General Staff College		
	Collecting and Utilizing Feedback: Combined Arms and Services Staff School	Ekwall, R.W.	943
	By-laws of the Military Testing Association		949
	Minutes of the Steering Committee		953
	Harry H. Greer Award		956
	MTA 27th Annual Conference Staff		957
	Military Testing Association Conference Registrants		958
	Author Index		970

PRE-COURSE MATHEMATICS SKILLS AND SUCCESS ON ADVANCED NAVAL TECHNICAL COURSES

Captain Kenneth A. Amyot

Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario, Canada

Background

Training for Naval trades in the Canadian Forces (CF) has become increasingly more expensive and lengthy in recent years. High levels of attrition on technically advanced weapons and electronics courses have become a cause for concern. To place the attrition problem into perspective, it is necessary to briefly describe the selection and training process.

Although Navy recruits are enrolled directly into electronics and weapons technician trades, their initial or basic training, called Trade Qualification (TQ) 3, is designed to prepare them to become junior operators of the naval combat systems equipment associated with their assigned trades. This is followed by two to three years of employment at sea during which time they qualify to the next higher level, TQ4, through on-the-job training. At this point, technical training to prepare them for employment as maintainers of equipment is provided as a TQ5 formal course.

In most CF trades, TQ5 training is a logical extension of TQ3 training, with success on basic (TQ3) training being a good predictor of success on more advanced (TQ5) training. For this reason, selection standards have been developed and validated on the basis of TQ3 performance. High success rates on Naval TQ3 courses indicate that the entry standards work well for that purpose. However, because of the disjuncture associated with the Naval "user/maintainer" concept, the aptitudes required for success at the higher levels may well be different than for the basic level.

Advanced Naval technician training is substantially academic, emphasizing mathematics. Aptitudes used for selection may include ability to learn, but not necessarily achievement; i.e., whether the material has already been acquired. Training development authorities at the CF Fleet School in Halifax, Nova Scotia, where these courses are conducted, attribute the high TQ5 attrition directly to a lack of sufficient mathematics skills prior to training. To address this deficiency, a three

---

The views and opinions expressed in this paper are those of the author, and not necessarily those of the Department of National Defence.

week Pre-Academic Qualifying Course (PAQC) was introduced to immediately precede the first semester of the TQ5 course. Although part of the purpose of the PAQC was to bring tradesmen up to the required threshold level of mathematics, in reality it serves a screening function. It has been increasingly difficult and expensive to continue to operate PAQC training. The Canadian Forces Personnel Applied Research Unit (CFPARU) was tasked to explore the possibility of a cost-effective substitute for PAQC training.

Remedial Training

Cognizant of the deficiency in pre-course mathematics skills, the CF Fleet School devised a number of remedial measures. A self-study Programmed Instructional Package (PIP) was found to be ineffective. A Computer Assisted Learning (CAL) Laboratory was established to allow students to upgrade their mathematical knowledge during off-duty time before or during TQ5 training. While effective, the CAL facility (located only on the East coast) was not readily accessible to all students. The scope of this project was, therefore, expanded to include an examination of means to better integrate the CAL facility within the training process.

Canadian Achievement Tests

Research was initiated to identify standardized, commercially available tests to measure competence in mathematics and to determine their ability to predict success on the PAQC and follow-on TQ5 courses. The tests selected were the two mathematics subtests of the Canadian Achievement Tests (CAT) (Canadian Test Centre, 1981). The California Achievement Tests, a widely used test battery in the United States, provided the initial pool of items for the CAT. The Canadian version met the criterion of being achievement tests for which Canadian national norms were available. The Canadian norms are based on a sample of 76,000 students randomly selected from across the country.

The first subtest, "Computation", relates to the four basic functions of addition, subtraction, multiplication and division. The second subtest, "Concepts and Applications", includes problems covering such areas as number theory, scales, measurement/graphs, geometry, problem solving, fractions, and rounding and estimating. The content of the subtests covers the full range of mathematics taught in grades nine through twelve. The subtest scores, and their combined total, may be used to obtain both norm-referenced and criterion-referenced descriptions of achievement.

Analysis

The two CAT subtests were administered to 166 Naval Electronics and Naval Weapons Technician students undergoing training at Canadian Forces Base (CFB) Halifax, Nova Scotia and CFB Esquimalt, British Columbia, during the first day of their PAQC training. A supplementary questionnaire, designed to obtain information on the students' mathematics background, was also administered. Pearson product-moment correlations were computed between each CAT subtest score and the total of these two tests, and the

results of the PAQC mathematics examinations and final overall course results. As well, grades obtained on mathematics examinations administered at the end of Semester I of the TQ5 course were correlated with CAT scores and PAQC results. Results in Table 1 indicate that the CAT is a valid predictor of success on PAQC and TQ training. Since the CAT predicts PAQC success, and since the PAQC predicts Semester I success, it was concluded that the PAQC could reasonably be eliminated in favour of CAT-testing.

Table 1

Inter-Correlations of CAT Scores,
PAQC and Semester I Performance Measures

	CA	Total	PAQC Math	PAQC Grade	Semester I Math
Comp	.76	.94	.75	.65	.50
CA	-	.93	.71	.61	.62
Total		-	.78	.69	.61
PAQC Math			-	.90	.70
PAQC Grade				-	.75

Notes: Sample sizes vary from 56 to 166 due to missing data.
All correlations are significant at the .001 level.

Reported Mathematics Education Versus CAT Scores

The correlation between grade equivalents measured by the CAT and grades levels reported on the supplementary questionnaire was low (r=.34). This correlation was calculated in several ways, controlling for other variables which might have affected performance such as academic upgrading since enrolment, use of the CAL Laboratory, academic stream (university preparatory or general/terminal), and type of program (business/ commercial, technical/vocational, or academic). This correlation was never higher than .40. Undoubtedly, other factors are involved, such as variability in inter-provincial standards (Ellis & Amyot, 1984), number of mathematics courses taken, and the interval between school leaving and CAT-testing. This points out the unreliability of grade levels reported at Recruiting Centres (even if accurate reporting is assumed) and the problems associated with using these as selection standards. To illustrate, several examinees reported a pre-testing mathematics education of grade nine or less, but their CAT grade

equivalents were in the grade twelve range. Similarly, several examinees who reported post-secondary mathematics at the time of enrolment scored less than grade nine equivalent on the CAT.

Diagnosis Versus Selection

Given the validity of the CAT for predicting success on PAQC and TQ training, the problem became one of when and how to use the CAT, since the primary purpose of the PAQC is screening, not teaching. One proposal was to use the CAT as a means of selecting recruits for Naval technician trades, by administering them during processing at CF Recruiting Centres. However, this was considered problematic for three reasons:

a. knowledge/proficiency measured at the time of enrolment can easily have deteriorated by the time the sailor reaches the point where he will proceed on the advanced training, which may be as long as three years later;

b. opportunities and programs exist for sailors deficient in mathematics to upgrade these skills between enrolment and commencement of advanced training; and,

c. testing time at Recruiting Centres is a major consideration, and the computation subtest alone requires at least twenty-eight minutes to administer.

This still left open the possibility of using the CAT at some point after enrolment to identify those students who are weak in mathematics. It was thought that the CAT might be administered, for example, during the preceding TQ course to identify potential failures on semester training, at which point a compulsory remedial program could be scheduled for these individuals.

Computer Assisted Learning Laboratory

Clearly, the existing Computer Assisted Learning (CAL) Laboratory, could be utilized to deliver the remedial program. The CAL Laboratory uses computer assisted instruction and a course authoring system developed jointly by the Ontario Institute for Studies in Education (OISE) (Gershman & Sakamoto, 1982) and Honeywell Information Systems (Honeywell, 1983) which can be conducted on many kinds of minicomputers. This particular system has been used successfully by 55 licencees in ten countries, including the Training and Development Branch of the Internal Revenue Service, Austin, Texas, who report a 60% reduction in training time compared to previous methods, thereby saving thousands of hours of student time with corresponding cost savings. There is a large collection of fully-validated courseware available including prerequisite, technical and academic mathematics. Most of the courses are tutorial in nature, intended to individualize instruction by diagnosing skill deficiencies and providing lessons as required. Techniques such as random test item generation, immediate answer analysis and feedback, and performance-dependent routing through a hierarchical structure of objectives, are used. Sea trials of the system (Shipboard Mathematical Refresher Training

(SMART)) have been conducted aboard a destroyer using a microcomputer. The CAL Laboratory is also used to provide computer-based instruction on advanced Marine Engineering training. However, as previously noted, participation in the remedial aspect of the CAL Laboratory has been on a voluntary basis.

Integration of CAT-Testing with CAL

CAT-testing is clearly a cost-effective alternative as a screening device, and has several additional advantages. The first is that it provides the potential to identify much more precisely the level and the specific skill/knowledge deficiencies of the TQ5 candidate than is possible with the PAQC. Second, it is simpler to assess a large number of candidates at various times and locations using a one hour test than it is to conduct a three-week course. This means that all TQ5 candidates can be assessed when and where they are available and at a point in time where an effective remedial program can be put in place. There is a considerable investment on the part of the CF and the individual during the years between enrolment and TQ5 training. The aim, therefore, should be to capitalize on this investment by bringing students to a point where they can succeed on this demanding, lengthy and expensive training. CAT-testing cannot, therefore, merely be substituted for the PAQC. There must be provision for remediation. While the CAL Laboratory has the potential to satisfy this requirement, it must be part of a coordinated program to be effective and efficient. The integration of CAT-testing and the CAL Laboratory should prepare students to pass the academic portions of TQ5 courses.

REFERENCES

Canadian Test Centre (1981). Canadian Achievement Tests. Scarborough, Ontario: McGraw-Hill Ryerson Ltd.

CAN-8 Instructional System: Summary description (1983). Willowdale, Ontario: Honeywell Information Systems Inc.

Ellis, R.T., & Amyot, K.A. (1984). Revised selection standards for re-structured naval trades (Technical Note 14/84). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Gershman, J.S., & Sakamoto, E.J. (1982). Computer assisted remediation and evaluation: Final report. Toronto, Ontario: Ontario Institute for Studies in Education.

Causes of Performance in Air Force Initial Skills Training

Francis D. Harding and Michael Mumford
Advanced Research Resources Organization

Joseph L. Weeks
Air Force Human Resources Laboratory

This paper reviews a study in which a system for assessing the impact of aptitude requirement adjustments on Air Force initial skills training was developed (Mumford, Weeks, Harding & Fleishman, 1985). As a result of the modeling process that was carried out, insights were gained into the initial skills training process to show how student attributes and course content variables interact to affect training outcomes.

In a previous but related research effort, the Air Force has developed an index of occupational learning difficulty defined as the time required to learn to satisfactorily perform the tasks required in an occupation. Subsequently, it has been suggested that the Air Force manpower allocation system might be improved if the Occupational Learning Difficulty Index was used as a frame of reference in establishing aptitude minimums. However, an analysis by Weeks (1984) indicated that the realignment of aptitude minimums in accordance with occupational learning difficulty would lead to marked shifts in aptitude minimums for several specialties. Thus, the concern for more definitive information about the possible impact of shifts in aptitude requirements before instituting such changes was expressed by Air Force managers.

Method

In reviewing the relationship between cognitive abilities and training outcomes, it became apparent that the effects of aptitude on training outcomes could be captured only by considering in addition to aptitude a variety of other student attributes as well as the potential interaction of aptitude and related variables with the training processes associated with particular training courses. Thus, it was decided that a multivariate path analysis approach would be the most appropriate means capable of incorporating a variety of student inputs, course content, and outcome variables.

The first step in the multivariate modeling effort was to define the major outcomes of initial skills training courses along with the student input and course content variables most likely to influence these outcomes. Also, at this time an attempt was made to identify potential measures of these variables and to identify any biases or limitations that might affect the general application of these measures. To obtain information needed to identify the potential variables, a series of structured interviews was conducted at several Air Force technical training centers. At each site variables indicative of the quality of student training performance; student input and traing content variables that might influence training performance; probable relationships between course content and student input variables; potential measures of student input, course content, and training outcome variables; potential sources and biases in these measures. As a result of

these meetings 7 student input, 16 course content, and 7 outcome variables were identified. The variables and their measures are shown in Table 1.

Table 1

Summary of Variables and Measures

Outcome Variables	Measure
Quality of Student Performance	Average of end of block tests
SIA Time	Number of hours of special individual assistance (SIA)
Academic Counseling	Number of academic counseling sessions
Nonacademic Counseling	Number of nonacademic counseling sessions
Retraining	Number of hours of retraining
Academic Attrition	Elimination for academic reasons
Nonacademic Attrition	Elimination for nonacademic reasons
Student Input/Variables	
Aptitude	Scores on selector aptitude index of ASVAB
Reading Level	Total score on the Air Force Reading Abilities Test
Academic Motivation	Number of difficult courses taken
Simple Interest	Received guaranteed specialty
Preference Interests	Received preferred specialty
Educational Level	Highest educational level attained
Educational Preparation	Recommended high school course prerequisites taken
Age	Years from birth
Course Content Variables	
Course Length	Total instructional hours in course
Day Length	Length of academic day
Student-Faculty Ration	Number of students per teacher
Instructor Quality	Average of instructor performance evaluations
Instructor Experience	Average months of instructor experience
Use of Aids	Number of instructional aids divided by course length
Hands-on Instruction	Hours of hands-on instruction divided by course length
Amount of Feedback	Number of evaluations divided by course length
Amount of Practice	Course length divided by number of its units
Reenlistment Bonus	Availability of selective reenlistment bonus
Yearly Flow	Number of students trained yearly
Occupational Difficulty	Overall learning difficulty of occupational tasks
Reading Difficulty	Reading grade level of course materials
Abstract Knowledge	Ratings of abstract knowledge required in course
Diversity	Number of units in course
Expected Attrition Rate	Expected attrition rate

Sample

Selection of courses to be included in the study was based on several criteria which reflected the pragmatic motivations underlying the study. Among the selection guidelines were that the courses should be representative of all Air Force initial skills courses; the four aptitude areas should be represented, there should be adequate variance in the minimum aptitude levels required for entry into the courses, high cost courses and courses with high student flow were to be chosen. The sample of trainees was chosen to include all who had entered the courses in the most recent six months

with a minimum of 50 students for each course. These procedures resulted in a total of 5,970 students in 48 courses. Thirty-eight of the courses and 5,081 of the students were used in the model development sample with the remaining courses and students held cross-validation sample.

Procedure

Once data collection and coding had been accomplished, preliminary data analyses were conducted in which it was assumed that all course content variables could be applied to all students in each course and that students with missing data for a variable would be omitted only from analyses which involved that variable. Descriptive statistics and bivariate correlations among all variables were obtained. This information provided a framework for reviewing the initial conceptual model and constructing a hypothetical model interrelating the student input, course content, and training outcome variables. This hypothetical model was used to generate the path specifications to be employed in a LISREL V analysis of the correlations obtained in the model development sample. The goodness of fit test and residual term generated by the LISREL V analysis were used to assess the extent to which the hypothetical model provided an adequate description of the interrelationships among the variables in the model. The multiple Rs generated by the model against each of the dependent variables were obtained along with the standardized path and regression coefficients generated in optimizing the fit between the hypothetical model and the observed correlation matrix. The regression and path coefficients provide a basis for assessing the causal impact of the student input and course content variables on training outcomes.

Results

Inspection of the bivariate correlations among all the variables which have import for understanding the causes of performance disclosed some interesting findings about Air Force initial skills training. First, the two interest measures produced a weak pattern of relationships which were hard to interpret so these measures were dropped from further analysis. The remainding student input variables, such as aptitude and student motivation, displayed positive relationships of about .50 and correlations with achievement test grades of between .20 and .35. Among the training outcome variables it was found that assessed quality of performance had negative relationships with the other training outcomes as would be expected given the impact of performance on counselling retraining and attrition. It was also found that counselling and attrition had weak relationships with student input and course content variables while assessed quality of performance and SIA time displayed moderate relationships with the input and course variables. This suggests that the effects of student and course attributes on outcomes are moderated through assessed quality of performance and SIA time.

Among the course content variables, instructor quality, instructor experience, amount of feedback, and course length were positively related to assessed quality of performance. Negative relationships with quality of performance were found for length of the academic day, student faculty ratio, student flow, and amount of practice.

Interpretation of the relationships among the course content variables led to the conclusion that occupational difficulty, manpower requirements and course subject matter difficulty were the prime determinants of the

nature of the training course. While occupational difficulty could be measured directly, manpower requirements and course subject matter difficulty represent latent variables that were defined through observed variables. As might be expected, manpower requirements were best defined by yearly flow, while course diversity, expected attrition rate, reading difficulty, and abstract knowledge requirements defined subject matter difficulty.

After the hypothetical model was refined on the basis of the obtained intercorrelations from the model development sample it was analyzed by LISREL V and found to provide an excellent fit to the observed relationships among the student input, course content, and training outcome variables within the model development sample as indicated by a goodness of fit index of .59 and residual of .19 that were obtained. A multiple R, training outcomes, and the predictive power of the model is shown by the .75 was obtained assessed quality of performance. Multiple correlation coefficients of .60 and .50 were obtained against academic and non-academically counselling respectively while retraining time, academic attrition and non-academic attrition produced Rs of .76, .83, and .59 respectively. The weakest prediction of an outcome variable was for SIA time which yielded an R of .35.

The standardized path coefficients generated by the LISREL V program are presented in Figure 1. In interpreting this schematic diagram of the model, the first of the three arrows entering the course parameter variables, e.g., course length, day length, student faculty ratio, etc., represents the effect of subject matter difficulty, the second arrow represents the effect of occupational difficulty, while the third arrow represents manpower requirements. Of the two arrows leaving the course parameter boxes, the first shows its impact on assessed quality of performance, the second on SIA time.

Inspection of the schematic diagram of the model indicates that the student input variables produced sizeable causal paths against assessed quality of performance: .16, .16, and .14 for aptitude, reading level, and academic motivation respectively. This suggests that other student input variables in addition to aptitude are prime determinants of training performance. Also, it should be noted that the prime course content variables, e.g., subject matter difficulty, occupational difficulty, and manpower requirements had negative effects on quality of student performance (-.08, -.13, -.10) respectively.

Assessed quality of student performance had substantial effects on all relevant training outcomes. Quality of performance had sizeable paths against academic counseling (-.20) and nonacademic counseling (-.23). Academic counseling in turn had a positive effect on academic attrition (.22) and on retraining (.34). These results show that counselling moderates the relationship between performance in training and subsequent outcomes such as retraining and elimination. This is one of the more interesting findings generated by the model, that is, the manner in which aptitude, or for that matter any student input variable, influences training outcomes. Aptitude and other student inputs have a strong direct causal impact on the assessed quality of student performance. However, the model indicates that student inputs do not have any direct effect on the more distal training outcomes such as academic attrition, nonacademic attrition, and retraining. Rather, student inputs affect the assessed quality of student performance

which in turn effects either academic counseling or nonacademic counseling. The fundamental importance of academic and nonacademic counseling in determining distal training outcomes is an important finding because it indicates that, regardless of the quality of student inputs or any manipulations made to them, whether an individual is eliminated from training or retrained is ultimately in the hands of course instructors. Given the fundamental importance of instructors' counseling decisions in determining retraining and attrition, it appears that more attention should be given to this process in future research efforts, and that both academic and nonacademic counseling should be considered major training outcomes. That retraining is an antecedent of academic attrition is shown by the path of .25. Thus, high retraining time appears to be an antecedent of attrition. Another interesting finding is that subject matter difficulty had a stronger causal effect (.16) on SIA time than did quality of performance (-.07). This suggests that poor performance might have some effect on SIA time, but that the difficulty of the course material appears to be the prime determinant.

While at first glance, the path coefficients obtained for many of the variables may seem low and appear to be lacking in effective predictive power, the sizes of the Rs generated by the model and the stable predictions generated in the cross validation courses shows that the model provides unusually effective predictions of training outcomes. The relatively low magnitude of the individual path coefficients really in effect confirms the initial premise of the study that technical training is a highly complex process which can only be described by a variety of student input, course content, and training outcome variables. While this is not an especially unusual conclusion, it suggests that optimal description and prediction of the training process can never be attained without employing multivariate techniques capable of taking into account this complex web of interrelationships.

References

Mumford, M.D., Weeks, J.L., Harding, F.D., & Fleishman, E.A. (1985). An Empirical System for Assessing the Impact of Aptitude Requirement Adjustments on Air Force Initial-Skills Training (AFHRL-TR-). Brooks AFB, TX: Manpower and personnel Division, Air Force Human Resources Laboratory. (In press)

Weeks, J.L. (1984, November). Occupational Learning Difficulty: A standard for determining the order of aptitude requirement minimums. (AFHRL-SR-84-26). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. (A417 410).

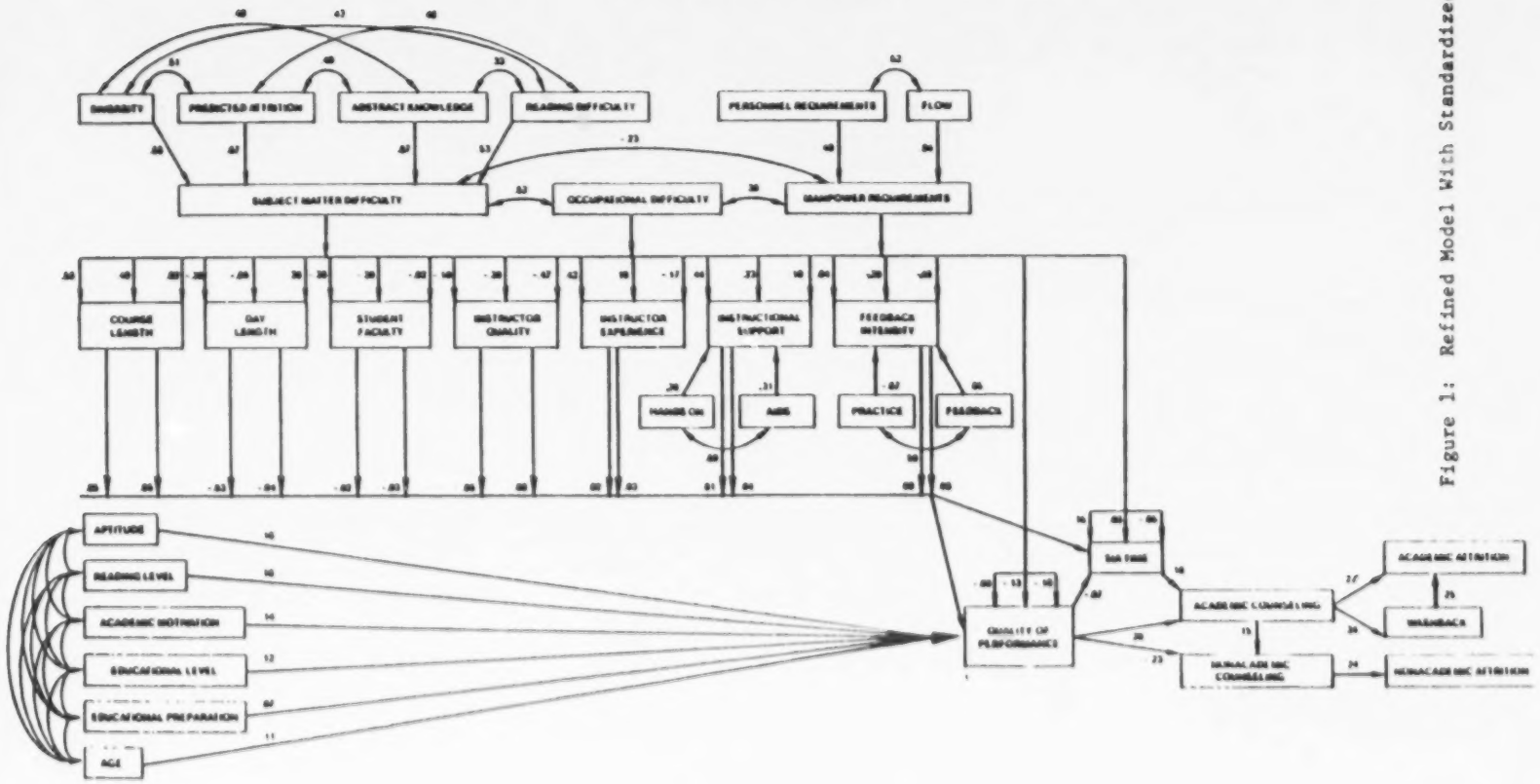


Figure 1: Refined Model With Standardized Path Coefficients.

AN EXAMINATION OF THE SIGNIFICANCE OF NON-PARTICIPATION IN FIELD STUDIES

Earl H. Potter III
Department of Economics and Management
U. S. Coast Guard Academy

Non-return of survey instruments is a problem that plagues all social scientists. In the analysis of returned surveys researchers generally try to demonstrate that there are no relevant differences between respondants and non-respondants. Most often this determination is made upon the basis of demographic data available to the researchers. Much other data that would be relevant and for which there could be differences is usually not available. This paper examines the personality differences between military participants and non-participants in three voluntary studies which required the keeping of event records over a six-week period.

Over the last several decades researchers have conducted numerous investigations to improve our understanding of the consequences for research validity of non-random attrition in field studies. Such investigations are usually limited by the non-availability of data on non-respondants, but significant differences have been found between subjects and non-respondants who can not be located for telephone interviews (Weaver, Holmes and Glenn, 1975), people who do not return mailed questionnaires (Macek and Miles, 1975) and persons likely to volunteer for experimental studies (Rosenthal and Rosnow, 1969). While the research literature is full of market research studies which examine non-return of questionnaires and the characteristics of volunteers, little work has been done on the characteristics of subjects who drop out of studies or renege on commitments to participate. Silverman (1964) has found that "no-shows" who had agreed to participate in a psychological experiment were higher in self esteem than were participants. Wrightsman (1966) reported that "no-shows" scored lower in social responsibility than did participants. Rosenthal and Rosnow (1969) concluded in a review of volunteerism that "no-shows" are similar to non-volunteers in nature and tend to differ from volunteers in several important respects. Most importantly, they noted that those conditions which support volunteering are exactly those which lead to subjects who had volunteered becoming "no-shows".

This is a significant point in considering research in the military under conditions where subjects are told that participation is voluntary, but hear that announcement from an official spokesperson in the chain of command. As researchers we hope that the suggestion that this is an official duty will encourage participation--dealing with non-participation is troublesome. But... subjects in fact have the right to withdraw. The past research suggests that these conditions may favor the selective withdrawal of certain subjects. This paper examines selective withdrawal in three studies at the U. S. Coast Guard Academy.

METHOD

SUBJECTS. The U. S. Coast Guard is the smallest of the U. S. Service Academies. Though less well known than West Point or Annapolis, the Coast Guard Academy functions much the same and in fact exchanges students with the other Academies regularly. Cadets enter the Academy at the beginning of the summer preceding their freshman year. The training that takes place during the summer is conducted chiefly by cadets (the Cadre) about to enter their own junior year. This period, called "Swab Summer", is, among other things, intended to be stressful and has changed little since Dornbusch (1955) used the Coast Guard Academy as a model for military socialization. During the remainder of a cadet's freshman year, the cadet is supervised closely and follows a strict academic, athletic and military regimen. While academic grades are given much as they are at any college, cadets are also rated for military career suitability by their peers.

STUDY DESIGN. Study I was planned as a field experiment in which one treatment group would receive a manipulation intended to focus a cadet's attention on his own ability to meet and deal successfully with the challenge of Swab Summer. A second treatment group received an additional manipulation intended to increase support networks within the platoon. It was hypothesized that subjects in the treatment groups would see the Swab Summer experience as being more manageable and less stressful and report themselves to be more ready to meet the challenges of the fall semester than a control group. Furthermore, it was hypothesized that the support network treatment would result in an increase in the quality and number of social supports.

Study II was intended as a replication of Study I. Study III was designed to follow up on the unexpected outcomes of Studies I and II. In Study III the nature of the coping skills treatment was reversed in order to assess the impact of a different instructional set upon perceptions of stress and performance.

PERSONALITY MEASURES. The Edwards Personal Preference Schedule (EPPS) was one of the personality measures. The EPPS is a forced choice inventory with items balanced for social desirability. The fifteen factors include achievement, affiliation, dominance and agression which have fairly clear associations as well as intraception, exhibition and order which generally are less familiar. The second scale was the California Personality Inventory (CPI). The CPI was designed to tap personality constructs with broad social relevance which favor the positive side of human nature. The eighteen subscales fall into four groups--measures of: (a) poise and self-assurance, (b) socialization and responsibility, (c) achievement potential and intellectual efficiency, and (d) interest modes. The third scale was the 16PF which was developed with normal and clinical groups. The sixteen independent factors of the scale are described with bipolar adjective sets in current handbooks; although, early in the development of the 16PF it was noted for the colorful, if obscure, factor labels chosen by the author. These factors include factor A (reserved vs warmhearted), Factor E (humble vs assertive), Factor L (trusting vs suspicious), Factor Q4 (relaxed vs tense). The three scales were administered to the entire entering classes early in their respective "Swab "Summers". In general, their profiles paralleled those of high school or college students with some notable similarities to military reference groups.

PARTICIPATION. The dependent variable is participation defined in Studies I and II as completing a stress diary for three of the six weeks of the study and in Study III as keeping a record of positive events for six of the six weeks of the study.

STUDY I

SUBJECTS. The subjects of Study I were the 351 cadets sworn into the Class of 1984 during the first week of July, 1980. Of these cadets 311 were men and 40 were women ranging in age from 17 to 21 with the majority (236) being 18 at the time of their entry. The subjects upon entry were randomly assigned by the Coast Guard Academy to one of nine platoons. The one exception to this procedure was that all the cadets who were band members were assigned to the same platoon. Since this fact resulted in the one platoon being systematically different both in composition and with respect to its program for SWAB summer, the band platoon was eliminated from this study as well as Studies II and III. The remaining eight platoons were randomly designated as belonging to Treatment A, Treatment B or Control groups. Platoons 1, 6 and 7 were assigned to Treatment A. During the summer program 12 of the original 117 cadets in Treatment A resigned. Dropping these cadets from the study resulted in Treatment A being comprised of 105 cadets in three separate platoons. Fifteen cadets resigned from the total of 114 cadets who started in Platoons 2, 4 and 8 which were designated as Treatment B leaving 99 cadets in Treatment B. Two platoons, with a total of 75 cadets completed the summer program in the Control group.

RESULTS. In Study I only four subjects in the social support network treatment and no subjects in the non-support treatment completed stress diaries for six weeks. Thirty-nine subjects met the participation criteria of three weeks. Participants were contrasted with non-participants using a T-test on the 16PF, CPI and EPPS. Of all the possible differences, only three emerged. On the 16PF participants appeared to be more conscientious (6.9 vs 6.2, $p<.05$) and more intelligent (6.9 vs 6.3, $p<.05$). On the CPI participants registered a lesser degree of flexibility (7.4 vs 9.0, $p<.01$) which Gough (1975) interprets as meaning that these subjects tend to be industrious, methodical, and responsive to authority.

STUDY II

Study II was conducted in part as an effort to improve the participation rate. The study was the same in all respects except that the experimenters increased their follow-up efforts by maintaining much closer contact with company officers and cadre than in the previous year. Despite these efforts, support for the study among company officers and cadre varied widely. Some responded to increased contact with more support for this study, some developed a vested interest in their own cadets' participation and supervised more closely than requested or desired and others flatly refused to encourage the participation of their cadet subordinates. What is noteworthy about this variability, and at the same time not unusual, was that command policy was supportive of, although not enthusiastic about, the study. Hence, all the

company officers and members of the cadre were publicly supportive and publicly committed to encouraging but not pressuring cadets to particicipate. The variability occurred in private statements and in behaviors contrary to public positions. Thus, while the experimenters in their introduction attempted to provide a more supportive and a less pressured introduction to the study, staff follow-up resulted in more pressure which probably contributed to a higher participation rate and, some cadets perceiving that the study was compulsory and not at all voluntary. This feature of Study II probably is common to most "voluntary" participation in officially sanctioned studies, but is clearly undesirable.

SUBJECTS. Study II was conducted during the summer of 1981 utilizing the 396 cadets of the Class of 1985 as subjects. Of these cadets 331 were men and 65 were women. They were divided into nine platoons one of which, the band platoon, was ommitted from the study. Treatment A, the non-meeting group, was comprised of three platoons totalling 133 cadets, 14 of whom resigned before the end of SWAB summer. The non-meeting group included therefore, 119 cadets. Treatment B, the meeting group was comprised of three platoons totalling 141 cadets of whom 15 resigned without completing the summer training program. The meeting group, therefore, included 126 cadets. A control group of two platoons with 85 cadets received no treatment as in Study I. Ages of the cadets were typical of cadet classes ranging from 17 to 21 with 61.6% being 18.

RESULTS. As in Study I, cadets in the meeting group did not appear to keep their committment to meet regularly in small groups. As did cadets in the non-meeting group, they discussed the study with friends of their own choosing. They seemed to form no significant bonds as a result of the randomly generated meeting groups. Given the tremendous pressure on cadet time it turned out to be next to impossible to get cadets to take the time to do something that was not natural when the experimenters could not require or even closely monitor compliance. These further findings support the determination made in Study I to consider participation in both Treatment A and Treatment B as members of one treatment group.* In Study II, 91 (out of a possible 245) cadets participated for three or more weeks of the study.

PARTICIPANT VS NON-PARTICIPANT DIFFERENCES. Given the greater participation rate and the perhaps different dynamics of Study II, one would not necessarily expect participants to differ from non-participants in the same fashion as in Study I. Both participants and non-participants rank above the normative means on the CPI's scales for dominance, capacity for status, and social presence. Non-participants, however, are characterized by a greater degree of dominance (3.9 vs 31.6, $p<.001$), greater capacity for status (20.5 vs 19.4, $p<.05$), and greater social presence (39.5 vs 37.7, $p<.05$). Non-participants are, therefore, more independent, self-seeking and less patient and deliberate. On the EPPS, non-participants appear as more dominant (34.0 vs 24.3, $p<.01$) and less verbally aggressive (12.8 vs 14.2, $p<.05$). The overall picture of the non-participants in contrast to the participants is that they are less likely to accept direction from outside, and more likely to invest themselves in demonstrating their independence. As Brehm (1966) would suggest, they are more likely to resist further imposition on their freedom by refusing to participate when the cost to their advancement is low.

STUDY III

OVERVIEW. Study III was conducted in the same fashion as Studies I and II with the one critical differance being that subjects were instructed to keep daily records of the good things that happened to them instead of stressful events. Since this task was both less demanding and more pleasant, participation for the minimum period was expected to be greater. Also, for the same reasons a greater degree of participation was considered essential. While the recording of one stressful event required writing a paragraph and took some thought, reporting that the cadet received a letter could be done in three words with little thought. By changing the structure of the treatment, therefore, the criterion for participation was also altered. A second consequence of the change in the treatment was that officer and cadet leadership could more easily encourage participation. Nevertheless, variability was still evident ranging from the exertion of too much pressure to less than enthusiastic follow-up.

SUBJECTS. Study III was conducted during the summer of 1982. Owing to a change in the Coast Guard requirements for Academy graduates the Class of 1986 which entered during that summer was considerably smaller than previous classes. A total of 250 cadets including 206 men and 44 women were sworn in during the first week of July. These cadets were randomly assigned to six platoons one of which, the band platoon, was omitted from the study. Two platoons totalling 87 people were designated as controls who received no treatment instructions. One platoon of 41 persons was assigned to Treatment A, the non-meeting group. Of these, 7 persons resigned before the end of the summer leaving 34 in Treatment A. Two platoons totalling 81 persons were assigned to Treatment B, the meeting group. Ages ranged from 17 to 21 with the majority being 18.

RESULTS. Participation in general was much higher as expected. This fact is due most likely to the much greater ease of completing the task. The entry of one event, e.g. "got a letter" would classify a cadet as participant for the week; yet, the real impact of participation might be much less. Therefore, for Study III the participation level was set as six weeks for all subjects. Defined in this manner, 27 of the 107 subjects were classified as participants. This figure places the participation rate of Study III at 25% and between the rate of Study I (19%) and Study II (37%). Once again cadets failed to meet regularly in their small groups. This fact is evident in the finding that social networks did not reflect the impact of assignment to small groups, from discussions with cadets, and from follow-up reports that groups seldom met. The sole virtue in attempting to get groups to meet for each of these studies was comparability. In order to make sense of the negative results associated with random assignment to small groups, however, one only has to recognize that cadets spend 24 hours per day working and living in close association with a peer group. Unlike a large university or even a small liberal arts college, the Coast Guard Academy spends a whole summer before the commencement of the Academic year building peer groups. It is apparent that cadets considered the small group assignments as superfluous. The majority reported a high degree of satisfaction with social supports ($\bar{X}$=3.3 on a 7 point scale with 1= "high satisfaction with support received at the Academy" and 23% reporting some dissatisfaction).

PARTICIPANTS VS NON-PARTICIPANT DIFFERENCES. Participants in Study III were characterized on the 16PF as being more conscientious (6.28 vs 5.24, $p<.01$) and more forthright (4.16 vs 5.12, $p<.05$) than non-participants. On the EPPS participants appeared to be higher in affiliation (18.74 vs 16.32, $p<.05$). This profile, as does the profile of participation in Study I, represents participants as more open and more likely to follow through on commitments.

DISCUSSION

It is clear that cadets who dropped out of these studies differ from cadets who met the participation criterion in the studies. Participation rates varied with the assigned task and with the energy with which follow-up efforts were made. Participation also varied with personality such that participants like the volunteers studied by Rosenthal and Rosow were more compliant than non-participants. The significance of these findings will vary with the nature of the study. In these studies personality variables on which subjects differ did not correlate with dependent variables. Were these studies of job satisfaction, leadership, motivation or work habits non-participation would not be so benign. In a military setting where subjects might feel compelled to "volunteer" resistance to pressure might well be a systematic bias to research results!

BIBLIOGRAPHY

BREHM, J.W. - A Theory of Psychological Reactance. New York: Academic Press, 1966.

DORNBUSCH, S.M. - The Military Academy as an Assimilating Institution, Social Forces, 1955, 33, 316-321.

GOUGH, H.G. - Manual for the California Psychological Inventory. Palo Alto, California: Consulting Psychologists Press, 1975.

MACEK, A.J. and G.H. Miles - IQ Score and Mailed Response, Journal of Applied Psychology, 1975, 60, 258-259.

ROSENTHAL, R. and R.L. Rosnow - Artifact in Behavioral Research. New York: Academic Press, 1969.

SILVERMAN, I. - Note on the Relationship of Self-esteem to Subject Self-Selection. Perceptual and Motor Skills, 1964, 19, 769-770.

WEAVER, C.N., S.L. Holmes and N.D. Glenn - Some Characteristics of Inaccessible Respondents in a Telephone Survey, Journal of Applied Psychology, 1975, 60, 260-262.

WRIGHTSMAN, L.S. - Predicting College Students' Participation in Required Psychology Experiments. American Psychologist, 1966, 21, 812-813.

ADVANCING SURVEY DESIGN THROUGH TECHNOLOGY

Linda M. Doherty
Navy Personnel Research and Development Center

Recent development in computer capabilities and communications are enabling attitudinal surveys to be designed, developed, administered, and analyzed quickly and accurately. The use of computerized survey systems should lead to improvements in survey design, while at the same time providing policy makers with timely and accurate information. This paper describes the components one such automated system, and outlines some important survey research issues that may be addressed by using this technology over traditional paper and pencil survey methods.

Computer technology applied to surveys in the past has focused almost exclusively on the statistical processing of data. While this is an important component of the survey research process, other components may now be automated and integrated into a complete system. Recent developments in computer capabilities have enabled the private sector to design, develop, administer and analyze surveys quickly in a broad range of settings that include local and national elections, and consumer product evaluation. Existing computer technology is sufficiently inexpensive so that computers can communicate with remote terminals to efficiently collect and analyze attitudinal information and integrate that information with other data bases. Since computer administration of surveys is a relatively new phenomenon, little research has been done comparing this method with other data collection methods. Also, survey design research issues more easily studied with this method have not yet been identified.

This paper focuses on (1) describing the components of a computerized survey system that should lead to improved survey design, and (2) defining some of the research issues that may be addressed through such automated data collection methods.

Automated Survey System Components

Automated Survey Administration and Data Collection

The Navy Personnel Research and Development Center is developing an automated survey system to assess the attitudes of the Navy civilian workforce. CENSUS (Computerized Executive Networking Survey System) administers questionnaires and collects information using remote terminals linked by phone lines to host computers. The attitudinal information is then integrated with existing personnel data bases and statistically analyzed.

Besides the efficiency of administering surveys via computer, the technology has other implications for survey design. First, the "demand" characteristics of the computer terminal are such that it is almost impossible to leave items blank, or provide more answers than requested. Second, since branching and individually tailored questions are easily presented, it is possible to study different versions of a survey upon individuals, and vary format characteristics, etc. Third, the survey format may be used to provide feedback to participants to increase the

accuracy and reliability of the information provided, and to present results from previous surveys. Feedback could include specific instructions about how to interact with the computer and complete the questionnaire, and provide positive and negative comments contingent on the quality of the responses.

Sampling Strategy

The sampling process may be made more efficient by computerizing the strategy used to extract samples from the population, permitting policy makers to select subsamples of interest, yet ensuring an appropriate mix of participants. The computer program will be structured so that an updated population data base may be maintained and monitored, enabling representative replacement participants to be selected efficiently. In addition, results from secondary subgroup analyses on the data base should well represent the population of interest. While the sampling strategy itself cannot directly influence the number of survey respondents, the computer terminal used to administer the surveys is interesting and involves the participants in the task. This should increase the likelihood that the identified sample will agree to respond to the survey.

Data Analysis

The automated survey system allows results to be computed more quickly and in greater depth than by transferring paper and pencil responses to computer, since the questionnaire is already in the appropriate format for data analysis. By automating the data collection and statistical analysis phases, data entry errors will be avoided. The faster the results are reported, the more timely and credible they will be, and, hence, more likely to be accepted and utilized by policy makers. Also, answers will be integrated into longitudinal data bases, where secondary analyses may be conducted with relative ease. Branching characteristics of the data collection methods will create hierarchically ordered data. Statistical methods need to be developed to analyze these type of data, and further advances in application of statistical and computational methods should occur as a result of the need to analyze longitudinal, hierarchically ordered data.

Computerized Authoring System

A computerized authoring system for survey development will be an interactive software program that provides a manager untrained in questionnaire design with a prompting, retrieval, and feedback system to produce reliable and useful surveys. As methods for developing questions become more precise, and as data banks of questions and responses are established, it is possible to develop an authoring system that can prompt managers to develop questionnaires independently and interactively. The authoring system will consist of two components - a menu-driven control system that includes an editor and format specifier, and a data base that includes survey data from previous questionnaire administrations and standardized questionnaires. The authoring system will improve questionnire design in general, and will enable the development of questionnaires from an **a priori** knowledge and theory building perspective. Thus, the quality of surveys should improve, as well as their applicability.

Feedback Systems to Policy Makers

The expected outcomes of automated surveys are results that are improved in quality and timeliness. By automating reports of the results, policy makers should be able to utilize the findings more quickly. Communication should improve by using the technology of presentation and communication methods. These feedback methods would include improved information displays such as the enhanced use of graphics (Fienberg, 1979). Also, a true feedback system will be established where policy makers become involved in developing questionnaires through the computerized authoring system and accessing the data base, interactively, through easy to use menu-driven computer programs. By evaluating the needs and questions of policy makers, more effective feedback that provides input to future survey designs is the final component in the automated survey system.

Research Issues

Response Effects

Most sources of variation in the quality of survey responses may be attributed to what Orne (1969) originally referred to as the "demand characteristics of the situation." Bradburn (1983) stated that "... the characteristics of the task are the major source of response effects and are, in general, much larger than effects due to interviewer or respondent characteristics." (p.291). Studies of the task characteristics in traditional paper and pencil questionnaires indicate there are few differences in the variability or quality of responses for different methods of presentation for all types of questions (Sudman and Bradburn, 1974). Use of the computer provides the survey developer with the ability to study these effects systematically.

Sensitive questions. Collecting responses to sensitive questions may be one area where method of presentation is important. In this area, computerized administration can make a considerable contribution by improving the number and quality of responses obtained. Respondents may perceive that completing a survey on an automated system is more anonymous than other methods, and, hence, they may be more willing to disclose personal and sensitive information. One method previously tested to increase anonymity, and believed to increase honesty, is the randomized response technique (Warner, 1965), which requires the respondent to use randomization to determine whether a sensitive question or an innocuous one is to be answered. Simple probability calculations then give an estimate of the nunber in the sample who agreed with the sensitive question, without knowledge of which respondents did so. This technique was found to increase reports of socially undesirable behavior and, perhaps, honesty. Some evidence indicates that respondents may answer electronic surveys with more extreme, thus, less socially desirable answers than they answer paper and pencil surveys (Kiesler and Sproull, 1985; Kiesler, et al., 1985).

Questionnaire format. Considerable work has been done to study the effects of task variables such as questionnaire format, wording, and length, with conclusions that these effects are too complex to make definitive statements (Sudman and Bradburn, 1980). Schuman and Presser (1981) found little relation between measures of attitude strength and

propensity to be influenced by question form. However, they did find that the context in which a question is asked - the ordering of questions, the inclusion of other questions and the arrangement of the questionnaire can produce response effects. The implications for studying response effects using an automated system is that it provides the flexibility to vary the form of the questionnaire, its length, wording, level of branching or contingency questions within a single questionnaire or among questionnaires. The Census bureau is testing one automated system, CATI (Computer Assisted Telephone Interviewing) system, where interviewers interact with a computer to make contact with respondents by telephone, and ask them branching, in-depth questions that are recorded immediately into the computer (Nicholls, 1983). With the flexibility in question presentation permitted by automated survey systems and sophisticated sampling plans, it is possible to design experiments that will elicit the specific characteristics of questionnaires to which respondents are most sensitive, and how those characteristics affect their answers.

Response sets and memory. Characteristics of respondents, or response sets, may influence answers to questions in a way that has nothing to do with the question being asked. Some research indicates (Bradburn et al., 1979) that response sets may not be respondents attempting to present a socially desirable image, but may reflect more stable personality characteristics or different life experiences. By exploiting the branching capabilities of the computer, it may be possible to discriminate between response sets and real differences in attitudes.

The effects of memory operate on responses because respondents forget what they have previously answered, or they telescope time and report events happening more recently than they actually occurred. With a longitudinal panel design for data collection and an automated survey system, the effects of memory may be tested. Previous responses and information about respondents may be resident in the survey software programs and used as memory aids when asking follow-on information. Also, updated information may be simply compared to previous responses.

Measurement Issues

One major issue in measurement theory focuses on the number of multiple items or indices required to adequately measure a psychological construct versus the number of questions that are reasonable to include in a survey of limited time (Anderson, et al., 1983). Advantages of a branching, computerized system are that the questions and response scales could be tailored to individuals, probing one issue in depth if appropriate. Scales could be utilized that would be sensitive to capture differences in individuals' attitudes and then using the flexibility of the system, adapt questions and scales to more precisely measure individuals attitudes. In this way a number of items would be used to assess a construct, but the number and type of questions and scales could vary.

Scaling techniques. Another implication for the data collection technology is the ability to more easily use complex scaling techniques to assess attitudes. While many of these techniques provide researchers with additional information, some of the shortcomings are that they are too cumbersome and time consuming (Anderson et al., 1983). For example, application of nonmetric multidimensional scaling techniques are

advantageous in the level of measurement required (ordinal) and output provided (dimensional plots), but the paired comparison method of data collection takes too much time and is uninteresting to the participant. With an interactive computer system, the data collection time could be reduced as additional data is requested only if other information is needed to construct an adequate perceptual map of the items.

Measurement models. A second major area in measurement that will be influenced by the applicability of technology to collect and analyze data is the use of more sophisticated measurement models. While methods to combine items to uncover a psychological construct are not new, development of these in the field of attitude measurement is. One technique, latent structure analysis is being developed in attitude measurement, and may provide a way to measure attitude types. (Reiser, 1981). The use of the broad set of multivariate techniques (models) should also increase. The appropriate use of factor analysis is one such important technique in describing attitudes.

Social Indicators

Automated surveys provide the ability to collect vast amounts of data rapidly and maintain large longitudinal data bases. The growth of these data bases should stimulate the growth of statistical and computation methods to analyze these specialized hierarchically ordered data bases, created because of the branching characteristics of automated data collection. In particular, the type and quality of the data should stimulate innovative methods of analyses by disciplines other than social science. Causal models that combine attitudinal and economic variables to describe behavior for individual subgroups of the samples should be one area that is stimulated by the accumulation of large data bases.

The second outcome of developing longitudinal data bases may be the improvement of experimental designs to evaluate the effects of policies. By integrating innovative experimental designs with survey research methodology, social indicators of the behavior of the military and civilian workforce may be developed. Collecting longitudinal, panel data in an efficient manner permits the establishment of baseline data and the ability to measure the long term effects of policy changes. Improvements to the design of experiments testing policy changes and the associated complicated sampling plans for collecting survey data may be the result of the implementation of automated surveys.

Conclusions

Automated survey systems will have a positive impact on survey design, and its methodological problems. Through computer technology, surveys will be able to be developed, administered, and analyzed more efficiently, providing high quality results to policy makers. Automated surveys can also substantially increase the amount of available data for subsequent analyses. One of the challenges will be to develop the analytic and statistical tools to evaluate the data, and to develop behavioral models using attitudinal and other, traditionally collected demographic variables. What is needed is the transfer of methodologies across academic disciplines to actively address the problems encountered in implementating automated survey systems.

References

Anderson, A. B., Basilevsky, A., & Hum, D. P. J. Measurement: theory and techniques. In P. H. Rossi, et al. (Eds.), Handbook of survey research. New York: Academic Press, 1983.

Bradburn, N. M. Response effects. In P. H. Rossi, et al. (Eds.), Handbook of survey research. New York: Academic Press, 1983.

Bradburn, N. M., Sudman, S., et al. Improving interview method and questionniare design: response effects in threatening questions in survey research. San Francisco: Jossey-Bass, 1979.

Fienberg, S. E. Graphical methods in statistics. The American Statistician, 1979, 33, 165-178.

Kiesler, S. & Sproull, L. S. Response effects in the electronic survey (Unpublished paper), 1985.

Kiesler, et al. Affect in computer-mediated communication: an experiment in synchronous terminal-to-terminal discussion. Human-Computer Interaction, 1985, 1, 77-103.

Nicholls, W. L., II. CATI research and development at the Census bureau. Sociological Methods & Research, 1983, 12, 191-197.

Orne, M.T. Demand characteristics and the concept of quasi controls. In R. Rosenthal and R. L. Rosnow (eds.), Artifact in behavioral research. New York: Academic Press, 1969.

Reiser, M. Latent trait modelling of attitude items. In G. W. Bohrnstedt and E. F. Borgatta (Eds.), Social Measurement: Current Issues. Beverly Hills: Sage Publications, 1981.

Shuman, S. & Presser, S. Questions and answers in attitude surveys: experiments on question form, wording and context. New York: Academic Press, 1981.

Sudman, S. and Bradburn, N. M. Response Effects in Surveys: A Review and Synthesis. Hawthorne, NY: Aldine, 1974.

Warner, S. L. Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Society, 1965, 60, 63-69.

IMPLEMENTATION OF A CIVILIAN AUTOMATED SURVEY SYSTEM

FRANK L. RIPKIN
STEVE CECIL

OFFICE OF THE CHIEF OF NAVAL OPERATIONS
CIVILIAN PERSONNEL POLICY DIVISION
ARLINGTON ANNEX
WASHINGTON, D.C. 20350

CENSUS -- The Computerized Executive Networking Survey System is being developed by the Navy Personnel Research and Development Center to address both an immediate operational need as well as to serve as a basis for new technology for automated surveys. This paper will be concerned with the concepts underlying the development of CENSUS and its potential application as a tool for policy development for the Navy.

The Department of the Navy has approximately 350 thousand civil service employees located worldwide, working in some 450 different occupations. With a work force of this magnitude, civilian policy makers require a considerable amount of information and data to formulate policies which range from procedures and regulations concerning the hiring of employees to policies governing their retirement system. The problem facing the policy maker is deriving information of both a factual (work force demographics, trends in retention, etc.) and an attitudinal nature in order to make possible proactive rather than reactive policy decisions. Compounding the problem is the requirement that policy be developed in a timely manner; we often have very little time to assess and respond to policy proposals presented to the Navy by such external agencies as the Office of Management and Budget or the Office of Personnel Management.

Using this as a background we normally have two alternatives for most policy decisions which impact on employees, their opinions and attitudes:

(1) The Quick Fix - - Providing for a gut or knee jerk reaction pertaining to employee attitudes coupled with an estimate of the demographic impact on the work force.

(2) The Steady Course - - Through this method we take the time, six to eighteen months normally, to develop quality survey instruments which address the issues in a straight-forward and accurate manner, identify a proper sampling that reflects the affected population and, conduct a thorough survey and analysis of the work force attitudes toward a given policy or policy proposal. We then couple this assessment of attitude and opinion with information detailing

work force demographic changes actually observed or modelled.

On reflection these two possible plans both offer a viable means for development of policy. The Quick Fix while fast, may result in a less than desirable outcome because a thorough analysis was not completed. The Steady Course on the other hand, offers a sound basis for policy development decisions, yet, by the time the decsison is formulated the issue has often been settled by an outside party (more than likely using the Quick Fix approach).

Thus the genesis of CENSUS - - Utilize advances in the technology of survey development, sampling techniques, telecommunications and computer analysis to reduce the time necessary to ascertain employee attitudes without sacrificing the accuracy of the Steady Course and while still gaining much of the speed of the Quick Fix. Joining these advances with those already being made by Navy in the area of Personnel data collection and analysis through the Navy Civilian Personnel Data System (NCPDS) and Human Resource Modelling, Navy Personnel policy makers will have a strong developmental tool.

The development of CENSUS has three objectives:

(1) General Objective: Develop a quick reaction automated survey system in support of Personnel policy development.

(2) Technological Objective: Develop sophisticated sampling plans.

Investigate the relationship between response effects and questionnaire/survey design.

Explore alternative survey presentations.

Develop causal models for application in policy analysis.

Develop a computer aided authoring system to assist managers in the development of survey instruments.

(3) Operational Objective: Assess the feasibility of collecting large samples of variables through an automated survey network.

Integration of up-to-date information on the attitudes and opinions of Navy employees with the information being obtained in advanced Navy data and modelling systems.

BEST COPY AVAILABLE

Development of longitudinal data bases to assess the long term impact of policies and the changes in employee attitudes over time.

CENSUS DEVELOPMENT

The development of the Computerized Executive Networking Survey System began in 1983 when the Office of the Chief of Naval Operations, Civilan Personnel Policy Division tasked the Navy Personnel Research and Development Center to develop a quick reaction survey system along the lines of the Harris or Gallup polls. The lead researcher on the CENSUS project is DR. Linda Doherty (please see a related paper elsewhere in this document). The project is to be developed in five stages:

(1) Initial systems test

(2) West Coast pilot implementation

(3) East Coast pilot implementation

(4) Nationwide expansion

(5) Implementation and ongoing research

The initial system and pilot tests for CENSUS were conducted in November 1984 and April 1985, in San Diego, California and in Washington, D.C., in August 1985. Approximately 1000 Navy employees have participated representing in excess of 30 Navy activities with usable results applicable to both the San Deigo and Washington areas.

SYSTEM DESIGN

The system components consisted of IBM PC AT & XT's with specially modified software, allowing simultaneous access by participants and presentation of the survey. The surveys were presented to the participants by means of a Northern Telecom DisplayPhone which served as both the telecommunications and terminal device. The micro computers served as host site computers during all tests. A mini computer was utilized as a secondary host during the initial system test. The West and East Coast test sites were established in order to test the reliability, speed, and utility of the host equipment, as well as the logistics of establishing regional survey centers. In order to make the survey test as real as possible actual surveys were developed by the staffs of NPRDC and this Office.

A unique aspect of the CENSUS surveys is that by presenting the survey via the computer and terminal we are able to eliminate two of the major drawbacks of a paper and pencil survey. One being the collection of background or demographic data on participants at the time of survey administration. All demographic data on

sampled employees is maintained at the host site with the participant's responses linked by use of their Social Security Account Number which is obtained during the sampling process and verified during survey administration. This speeds the survey administration as well as allows analysis of responses keyed to a large amount of demographic data, some of which may be added after the survey has been administered. This is something which is not normally possible in a paper and pencil survey, and can be a strong tool for analysis when twists and turns in policy development require additional or unanticipated information be supplied to the functional manager.

The second benefit of the computer aided survey process is the application of branching in the survey design. By making the process of branching transparent to the participant CENSUS has eliminated the standard and frustrating, "if your answer is yes please proceed to question number 34." We anticipate that the use of branching will make sophisticated survey design a routine matter in CENSUS and also allow immediate feedback of information to participants possible during survey administration.

ASSESSMENT

The issues under review during the system and pilot tests were the equipment and how well it would function, the software operations which allow multiple access to the host computer and stores the survey files, the terminal operations and its ease of use, the logistics of establishing the survey network, and the attitudes of the participants. Additional concerns involved the actual survey development process and the utility of survey results for policy development.

These tests have proven to be very satisfactory. The results showed the reliability of the IBM PC AT/XT as a host computer. It was determined, however, that even though the PC proved to be reliable, a back-up computer would be required in future sessions. The mini computer functioned adequately yet due to its age and mechanical condition did not prove to be as reliable as the micro. It was decided that further tests of a mini computer as a host would be delayed until a later date. The software functioned as planned, though results of the later survey pilots have shown improvements in file handling are required before full implementation would be possible.

The main concern from these system and pilot tests was the extensive logistics and up front work required to establish a survey site. The process is labor intensive, requiring personal contacts with activities and participants to be conducted before, during and after the survey. We anticipate the design of the implementation plan can take this into consideration and thus reduce the manpower and logistics involved. The negative aspects of manpower and logistics requirements will continue to be addressed as future pilots are conducted.

BEST COPY AVAILABLE

The most encouraging outcome from our prospective was the ease of use for the participants and the utility to policy development of the total CENSUS process. The participants found the survey easy to take, the branching of questions, which is transparent to the user, added significantly to the user acceptance. The general response of the user was one of enthusiasm.

The functional managers have found that the speed of development and the amount of specific data made available to them far exceeds previous survey processes. Following the analyses of the test surveys, data were presented to functional managers for evaluation as a tool for policy development. The results of the survey, while valid only for the area in which the test took place, contained enough specificity and information to rival previous paper and pencil surveys. As stated earlier previous surveys of this nature have taken 6 to 18 months to produce and analyze. Each of these tests took less than 3 months to develop. conduct and analyze. We anticipate that after full implementation of the system, CENSUS will be able to reduce the time for survey development and administration even further.

The functional managers also found the use of electronic mail between NPRDC and this Office during the second West Coast and the East Coast pilots to be extremly beneficial in the development of the survey instrument.

From analyses conducted by NPRDC and this Office, the initial system and pilot tests were determined to be a complete success. Further development of the CENSUS program will be continued concentrating now on system improvements, Navy-wide implementation planning, and continuing research in the areas of survey development and analysis.

Questions concerning the development and application of CENSUS may be directed to Mr. Steve Cecil, CENSUS Project Director. Mr. Cecil may be reached at the above address or by calling (202) 694-5762 or autovan 224-5762.

The Utility of Educational and Biographical Information for Predicting Military Attrition

Barry J. Riegelhaupt
Thomas P. Bonczar
Human Resources Research Organization

> An examination of discriminating item responses can tell a great deal about what kinds of employees remain on a job and what kinds do not, what kinds sell much insurance and what kinds sell little, or what kinds are promoted slowly and what kinds are promoted rapidly. Insights obtained in this fashion may serve anyone from the initial interviewer to the manager who formulates personnel policy. (Owens, 1976, p. 612)

Compelling evidence demonstrates that, when appropriate procedures are followed, the accuracy of biographical data as predictors of future work behavior may be superior to any known alternative (Cascio, 1982). Biographical inventories have been developed for a diversity of occupations, from unskilled and clerical workers to research scientists and Army officers. By far, the most commonly used criterion for developing, validating and cross-validating these inventories has been attrition.

The role that biographical information can play in predicting attrition, as well as job performance, both costly problems for the Military Services, has been recognized by the Department of Defense. The Office of the Assistant Secretary of Defense (Force Management and Personnel) as part of a study to validate and improve current education and moral enlistment standards, awarded a contract to the Human Resources Research Organization (HumRRO) to develop the Educational and Biographical Information Survey (EBIS). Analyses relating EBIS responses to military performance are continuing under a contract sponsored by the Office of Naval Research. The long-range objective of this research is to help the Services identify potentially successful nongraduates and refine selection among high school graduates.

EBIS Composition

EBIS items can be classified into nine constructs, theoretically related to military performance. Demographic items address type of area in which one grew up, family income, and parents' education. Education achievement variables deal with years of education, school grades, and type of education credential. School behavior/attitude items concern school activities, disciplinary incidents, reasons for thinking about quitting school, and so

---

This research was performed under contract to the Office of the Assistant Secretary of Defense (Force Management and Personnel) and the Office of Naval Research. The views and opinions expressed in this paper are those of the authors and should not be taken as official DoD policy.

on. Family relations items deal with attitude toward parental discipline, presence of each parent in the home, and parent stability (e.g., incidence of alcoholism, parent's arrest, etc.). Length of longest full- and part-time job and all the reasons the respondent ever quit a job are included in the work history category. Status variables are subject characteristics such as sex, year of birth, and kind of high school attended. Arrest-related items deal with traffic, misdemeanor, and felony arrests and convictions. Alcohol/drug use items deal with both frequency and age of onset for use of these substances. Minor Misbehaviors deal with youthful behaviors such as smoking or running away from home, which are not illegal, but might be construed as midly negative.

EBIS Administration

The EBIS was administered to approximately 75,000 military applicants and recruits in the Spring of 1983. Applicants took the EBIS along with the Armed Services Vocational Aptitude Battery (ASVAB) at a Military Entrance Processing Station (MEPS) or associated test site. Recruits were given the EBIS during in-processing at their Recruit Training Center (RTC). Individuals who took the EBIS are being tracked through their first terms of service, with an emphasis on identifying military suitability predictors within education groups. Based on simple correlations between EBIS items and attrition status, previous work indentified the 24 best EBIS items for predicting six-month attrition among high school graduates and the best 24 items for nongraduates (Means & Laurence, 1984). Building on that earlier effort, the present work describes the statistical weighting of EBIS items based on 12-month attrition status to produce a total EBIS score, and scores for each of the nine constructs. The utility of the EBIS for predicting 12-month attrition is examined. Of the approximately 75,000 EBIS respondents, only the 3,092 male nongraduates who took the EBIS as recruits were examined in the present effort.

Research Design

In addition to nonattrition (stayers) and attrition (leavers) groups for use in weighting biographical items to predict the criterion of 12-month attrition status, nonattrition and attrition hold-out groups from the same population are needed for evaluation of items and weights. Thus, in addition to dividing the sample into stayers and leavers, these two groups were further subdivided into weighting/selection and evaluation (hold-out) groups. For the sample of 3,092, the design is shown below.

	Criterion Group		
	Stayers	Leavers	Total
Weighting/Selection	1262	592	1854
Evaluation	842	396	1238
Total	2104	988	3092

Figure 1. Allocation of Subjects for Weighting and Evaluation

Item Weighting and Selection

While a number of different techniques are available for statistically weighting items (c.f. Owens, 1976), perhaps the simplest technique is the Horizontal Percent Method. Cascio (1982) notes that as long as samples are relatively large, this method yields results comparable to any of the more sophisticated procedures. An example of how items were weighted is shown in Figure 2. The frequency of respondents falling into each criterion group (stayers vs. leavers), response category combination was computed. Then, after adjusting for differences in sample sizes in the two groups, for each response category, (working horizontally) the percentage of cases in the nonattrition (i.e., stayers) group was computed. This percentage became the response category weight.

EBIS Item	Freq. of Stayers	Freq. of Leavers	Total Number	Percent Stayers
13. Were you ever expelled from school?				
Yes	140	107	247	.38
No	1086	471	1557	.52
	1226	578	1804	
28. Frequency of Physical Fights?				
Never	600	272	872	.52
Once or Twice	187	105	292	.47
Occasionally	30	23	53	.40
Fairly Often	12	15	27	.29
	829	415	1244	

Figure 2. Horizontal Percent Method of Weighting EBIS Items

This approach to weighting has a number of advantages. It can be applied, without modification, to all items regardless of the number of response options that exist for an item. Each option receives a weight. This is of particular importance for the EBIS where the number of response options range from two to thirteen across the 119 items. Since this approach does not require that all items have an equal number of options, it is not necessary to artifically dichotomize, for instance, all items for the sake of consistency in format. Finally, the computation of weights yields information relevant to the determination of item quality (i.e., response frequencies).

Thus, after the weighting algorithm had been applied to the weighting sample, response option weights and frequencies were examined for purposes of eliminating poor items. A poor item was defined as one that either did not discriminate stayers from leavers (i.e., response choices had weights around .50) or had little variance (i.e., greater than 85 percent of the respondents

selected one response option). The distribution of items retained for further analyses is shown in Table 1. As can be seen, the Demographic, Status, and Arrest Related scales were deleted.

Table 1

Distribution of Item Types Represented on the EBIS and Retained for Validation Against Attrition

EBIS Scale	# Items	# Retained	% Retained
Demographic	5	0	0
Education Achievement	14	7	50
School Behavior/Attitudes	32	16	50
Family Relations	13	2	15
Work History	17	5	29
Status Variables	4	0	0
Arrest Related	16	0	0
Alcohol/Drug Use	12	4	33
Minor Misbehaviors	6	4	67
Total	119	38	32

EBIS Scoring

Having selected 38 discriminating items, the next step was to compute an EBIS score for each individual in the item weighting/selection sample. The score an individual received for an item was the weight associated with the response option selected. For example, referring to Figure 2, if the respondent was expelled from school, a .38 would be the score for that item. If the individual fairly often engaged in physical fights, a .29 would be that individual's score for that item. A total scale score was computed for each individual by summing the response weights across the 38 items. In a similar fashion, individual scale scores were also computed.

Comparison of Stayers and Leavers

Following the computation of EBIS scores for individuals in the item weighting sample, stayers and leavers were compared on the total EBIS scale as well as the six factors that were retained. The results are presented in Table 2.

As can be seen, stayers and leavers had significantly different scores on the total EBIS scale, as well as five of the six factors. Only for Minor Misbehaviors was the mean difference not statistically significant, although it was in the expected direction. These results show that those who remain in the Military Services have different pre-enlistment experiences than those who leave, thus demonstrating the usefulness of a biographical inventory.

Table 2

Comparison of Stayers and Leavers in the Item Weighting/Selection Sample

	Stayers[a]		Leavers[b]		
EBIS Scale	Mean	SD	Mean	SD	t-value
Total	18.408	1.058	18.136	1.305	4.78*
Education Achievement	3.409	.422	3.334	.454	3.29*
School Behavior/Attitudes	7.897	.353	7.812	.401	4.65*
Family Relations	.979	.114	.965	.127	2.49*
Work History	2.513	.055	2.502	.056	3.97*
Alcohol/Drug Use	1.431	.242	1.398	.262	2.56*
Minor Misbehaviors	1.784	.309	1.763	.329	1.34

[a]n = 1262
[b]n = 592
*p <.05.

These results should be interpreted with caution, however. Since these comparisons were performed on the same individuals that were used to weight and select the items, observed differences between stayers and leavers may not reflect true differences, but rather chance fluctuations. The application of the 38 items with their corresponding weights to an independent sample might yield substantially different results from those obtained using the item weighting/selection sample. Thus, the next step was to evaluate the utility of the EBIS in the hold-out sample.

Comparison of Stayers and Leavers in Hold-Out Sample

As in the previous sample, item weights were applied to individuals in the hold-out sample and EBIS scores were derived. Once again, stayers and leavers were compared on a total EBIS score and on the six factor scores. The results are presented in Table 3. As can be seen, the EBIS total score, as well as the Education Achievement, School Behavior/Attitudes, and

Table 3

Comparison of Stayers and Leavers in the Evaluation (Hold-out) Sample

	Stayers[a]		Leavers[b]		
EBIS Scale	Mean	SD	Mean	SD	t-value
Total	18.384	1.029	18.195	1.266	2.79*
Education Achievement	3.410	.370	3.351	.474	3.36*
School Behavior/Attitudes	7.890	.328	7.821	.395	3.22*
Family Relations	.976	.121	.962	.133	1.32
Work History	2.511	.056	2.509	.057	<1
Alcohol/Drug Use	1.440	.211	1.408	.272	2.22*
Minor Misbehaviors	1.768	.330	1.786	.308	<1

[a]n = 342
[b]n = 396
*p <.05.

Alcohol/Drug Use scales yielded significantly different mean scores for stayers and leavers. The significant differences that were found on the Family Relations and Work History scales in the weighting sample, were not found in the hold-out sample. As in the weighting sample, scores on the Minor Misbehaviors factor did not differ when comparing stayers and leavers in the hold-out sample.

Conclusions

The data in this report demonstrate the utility of biodata for predicting attrition for nongraduate male recruits. Specifically, items dealing with School Behavior and Attitudes, Education Achievement, and Alcohol and Drug Use all showed significant relationships with attrition. While statistically significant, mean scale differences between stayers and leavers were relatively small. The conclusion, however, that these significant findings are of little practical value is unwarranted. As scaled, item weights can range from 0.0 to 1.0. A value of .50 reflects a nondiscriminating item response. As values move closer to 0.0 or 1.0, the response becomes a better discriminator. It is rare, however, to find response choices with sufficient numbers of respondents with weights approaching 0.0 or 1.0. If such items existed, they would likely be more predictive of attrition than most entire biodata scales. No such items exist on the EBIS. Typically, response choices within items differed by .10 or less. Thus, summing across items for stayers and leavers, produces relatively small mean differences. However, very small standard deviations also resulted, which contributed to the significant relationships found in the hold-out sample. It is the finding of significance in a hold-out sample that is the true test of an instrument's usefulness. The EBIS passed this test. Thus, biodata items, such as those found on the EBIS, may allow the Services to enlist nongraduates who have a better chance of fulfilling their enlistment obligation.

Next Steps

Future work will look at military applicants' EBIS scores to determine and adjust for the restriction of range that might exist in the sample examined in this paper. Additionally, EBIS cut off scores will be set and the amount of improvement in predicting attrition attributable to biodata will be examined.

References

Cascio, W. F. (1982). Applied psychology in personnel management (2nd ed.). Reston, VA.

Means, B. & Laurence, J. H. (1984, November). Improving the prediction of military suitability through educational and biographical information. In Proceedings of the 25th Annual Conference of the Military Testing Association, 203-208.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

THREE VARIABLES THAT MAY INFLUENCE THE VALIDITY OF BIODATA[1]

Clinton B. Walker
U.S. Army Research Institute for the Behavioral and Social Sciences

This research examines the effect on predictive validity of traditional procedures for developing and implementing suitability screens in the military. For this paper, suitability screens in the form of background questionnaires, or biodata, will be considered. Typically, predictor tryouts have been run on new recruits whose subsequent performance has been tracked for the first six months of service (Atwater & Abrahams, 1983; Walker, 1985). Item selection and keying have then been based on the observed relation between predictor data and the criterion of successful service (versus discharge for bad causes). In the case of the U.S. Army's Military Applicant Profile (MAP), the instruments and keys have been implemented no less than two and a half years after the tryouts.

There is reason to suspect that three aspects of this traditional sequence - viz., testing recruits rather than applicants, tracking the cases for only six months, and implementing long after pilot testing - adversely affect operational validities. Since recruits and applicants are likely to differ in their desire to make themselves look good on self-report measures, applicants could be expected to try more than recruits to earn high scores. As a result, scoring keys that are developed on data from recruits may be less valid for scoring responses of applicants. In support of this hypothesis, Means and Heisey (1985) have found more self-serving responses in data from applicants than from recruits.

The hypothesis that using only a six-month tenure for tracking success/attrition lowers validities is based on the following two premises. First, more than half of attritions occur after the initial six months (Goodstadt & Yedlin, 1980; Hicks, 1981; Walker, 1985). Second, attrition during the first six months may not occur for the same reasons as later attrition. In the first six months recruits make their initial adjustment to military life while undergoing entry-level training; after that they are serving with operational units. Unfortunately, the archival codes for types of attrition are too cryptic (e.g., "Trainee Discharge Program," "Unsuitable Unknown," "In Lieu of Court Martial") to indicate whether earlier and later attrition are qualitatively different phenomena. But if they are, then using a longer than traditional criterion period for developing scoring keys might produce different keys.

A long lag time before implementing scoring keys is suspect because characteristics of the applicant pool change over time. Once the predictor data are collected for developing a biodata instrument, they may obsolesce as the criterion ripens. If the nature of the applicant pool changes much, then a scoring key may lose validity before it is ever used for screening, and continue to lose validity after implementation.

[1]Thanks go to Elizabeth P. Smith for advice on programming and to Winnie Young for creating the dataset on applicants in FY 81/82 from the Project A Longitudinal Research Database. The views in this paper are those of the author and do not necessarily reflect the official positions or policies of the U.S. Army Research Institute or Department of the Army.

The present research uses data from the MAP to test the effects of each of these variables. Responses to a common set of items by contemporaneous applicants and recruits are compared to test the effect of examinees' status. Then, various statistics are examined over the course of years to test the effect of time lapse on the keys' validity. Finally, the effect of duration of the criterion period is tested by comparing the predictor responses of examinees who were discharged within and beyond the first six months of service. Each of those issues is treated in turn below in a separate section.

Applicants Versus Recruits

Method

To keep from confounding the effects of examinees' status with those of date of testing (i.e., temporal drift), it is necessary to compare contemporaneous applicants and recruits. Two such comparisons are available in the MAP data. First, MAP scores of 2,374 non-graduate applicants during FY 82 were compared with those of 1,286 non-graduate recruits who were tested in February-June, 1982. These recruits were the non-graduate subset of a sample of 9,603 cases on whom new instruments were being developed (Erwin, 1985). Out of the 240 items in that research, 38 were chosen for use here according to these two criteria: they had to be on the operational form of MAP, so the applicants would have taken them, and they must have shown validity for non-graduates in the developmental research. These 38 were the universe of items that met both criteria. The key for scoring had been developed on all 9,603 cases. Here the comparison was a t-test on the total score, 0 to 71 being the possible range.

Data for the second comparison overlap in part with the previous ones. In the developmental work of 1982, the item pool was administered to a sample of applicants at 39 Military Entrance Processing Stations (MEPS) nationwide and to recruits at all seven Army Reception Stations. Out of those groups, a respective 949 and 9,603 examinees of all levels of education, age, and gender were retained for analysis. Retention was based solely on the availability of individuals' criterion data in central personnel files. In the applicant sample, 267 cases retook the instrument later as members of the recruit sample. Presumably the presence of those cases reduces the between-group differences, thus biasing any test against finding differences.

The vehicle for this second comparison was two 101-item forms of MAP which were developed on the 9,603 recruits. These forms each had 78 unique items and 23 items in common, yielding possible scores of 1 to 188 on one and 0 to 194 on the other. Mean MAP scores and validities against the six-month tenure criterion were compared in the applicants and recruits.

Results

Descriptive statistics for the non-graduate applicants in FY 82 and the recruits in the 1982 development sample are included in Table 1. The observed difference in means of 10.9 points is significant (t = 52.9; $p < .001$) and the effect is strong (omega square = .43). Data for applicants and recruits in the 1982 developmental project are summarized in Table 2. Applicants' total scores were higher by 2.81 points on Form 1 and 2.1 points on Form 2. These differences gave t's of 5.62 and 3.96 ($p < .01$ in each case). However, here the strength of effect was less than 1% for each form. For both forms, the

observed validities were higher for recruits than for applicants. The difference between correlation coefficients for independent groups (Guilford & Fruchter, 1973) was computed on the validities for each form. The observed z's of 2.51 and 1.59 had one-tailed probabilities of .006 and .056, respectively.

Table 1
Descriptive statistics for four samples of non-graduates[a]

Logical role	Date of predictor data	Status: Applic/ Recruit	n	r	Mean out of 0-71	SD	% finish 6 mo
Develop key	1-6/82	Recr	1,286	.18	33.3	6.2	79
X-validation & 0 yr drift	10/81-9/82	Appl	2,374	.02	44.2	5.5	86
1 yr drift	10/80-9/81	Appl	3,567	.07	44.2	5.2	86
2 yr drift	7/79-6/80	Appl	14,771	.01	28.3	5.6	86

[a]The "instrument" for these data was 38 items from MAP 4B which were keyed on the total 1982 development sample of 9,603 cases and were also valid for its non-graduate subsample.

Table 2
Descriptive data for applicants and recruits in 1982 development sample

		Form 1			Form 2		
Status	n	Mean	SD	r	Mean	SD	r
Applicants	949[b]	125.56	14.49	.24	123.72	15.54	.27
Recruits	9,603[b]	122.75	16.64	.32	121.62	17.27	.32

[b]Both samples include 267 cases who took the instrument a second time as members of the recruit sample.

Discussion

Both sets of comparisons support the hypotheses that applicants get significantly higher scores than recruits, even though both samples were selected on the basis of operational MAP. Although the comparison of validities favors the hypothesis, that evidence is weakened by the fact that the recruit sample was also the sample on which the scoring key was developed. Nevertheless, the generalizability of data from recruits to applicants is not supported here.

Drift in Validity

Method

For examining possible loss of validity over time, a non-operational key was used that had been developed on the 1982 recruit data. The criterion was

successful completion of the first six months of service (vs. discharge for failures to adapt). The "instrument" consisted of the 38 items mentioned earlier. Meeting the criteria of being on the operational form of MAP and being validated on non-graduates, the items could be used to compare results for non-graduates in different year groups who took MAP before entering the service. Three samples of such applicants were available: 2,374 in FY 82, 3,567 in FY 81, and 14,771 in 7/79-6/80. Because the 1982 key was not cross-validated by the developer, the 1982 applicants became a cross-validation sample. Thus, their data were used to see how much validity there was to drift in the first place. Validities in the form of Pearson r's, mean total scores for the 38 items, and success rates (i.e., percent of sample completing the first six months of service) were compared over the four samples.

Results

Table 1 gives descriptive statistics for the recruits in 1982 and for three samples of applicants. In contrast with the original value of .18, validities for applicants in 1982, 1981, and 1979/80 were .02, .07, and .01, in order. The key did not effectively discriminate between examinees who went on to complete the first six months of service and those who did not: mean differences in scores for those two criterion groups reached a maximum of .18 SD in the three samples. Means out of a possible 71 points ranged from 28.3 to 44.2 points in the four groups, while success rates varied from .79 to .86. Using the 1982 applicants as a basis for confidence intervals on the means, we find significant differences ($p < .001$) in both the 1979/80 applicants and in the 1982 development sample. The normal approximation to the binomial found the development sample to have a significantly lower attrition rate than the 1982 applicants ($z = 12.28$; $p < .001$), all of whom had entered the Army.

Discussion

The low validity that was observed in the 1982 applicants amounts to a failure of the (non-operational) 1982 key to cross-validate. Thus, there was little if any original validity that could drift. Absent drift in validity, however, there were significant jumps in both predictor and criterion scores across samples. If changes occur in validity over time, they could be due to gradual trends in the population of applicants, to short range instability in the population, or to both. It is possible that similar variability could be found in subsamples of the 1982 recruits. In order for the 1982 developmental data to have any hope of producing a durable key, they would have to undergo a legitimate cross-validation. Elizabeth P. Smith and I are now working on this problem in-house at the Army Research Institute.

Six Months Versus Longer Tenure

Method

An operational form of MAP, Form 4B, gave the data for this analysis. Its 60 multiple choice items were validated in 1977 on 2,280 male recruits who had not completed high school (Frank & Erwin, 1978). In content, the questions cover experiences in school, extracurricular activities, work history, and expectations of life in the service. The present examinees were 2,564 17-year old non-graduate males. They all took MAP as a pre-induction screen in Fiscal Years 81/82, entered the Army, and then received adverse discharges in their

first tour. For the first analysis, examinees were split into two groups, those discharged within the first six months of service (n=860) and those with longer tenures ($\bar{x}$=363 days; n=1,704). For each of the 60 items, a chi-square test of association was run on frequencies of response for each alternative by group. Cramer's V for the items was examined as well for estimates of strength of effects. To judge the potential of response choices for keying, differences between groups in rates of endorsing individual choices were examined in items giving a significant groups-by-response choice chi square.

A second similar analysis was done to see whether the sensitivity of biodata items to individual differences in adaptability is masked by lumping successful cases with those who receive bad discharges after six months. For this analysis, chi-square tests were run twice on the total sample of 5,941 non-graduate applicants in FY 81/82. This sample included those who served successfully. The sample was split differently for these runs: once as all dischargees vs. all successful cases, and once as all discharges within six months of entry vs. all other cases. Simple numbers of significant ($p < .05$) chi squares and median p values from the two splits were compared.

Results

In the analysis of dischargees, 10 of the 60 group-by-response choice chi square tests gave probabilities < .05. Of those, three had p's < .01. Cramer's V for the ten items ranged from .056 to .085, while V's for seven items with .05 < p < .15 were also above .05. The median level of significance for all 60 items was .30. In each of the ten items with the lowest p values, the single response choice which had the greatest difference between groups in rate of endorsement was tallied. The median of those ten maximal differences was 4.2% (range: 3.19 - 6.78%).

In the second analysis, 13 of the chi-squares on items gave $p < .01$ when the positive criterion group included bad discharges after six months. In contrast, when the criteron groups are pure (i.e., all bad discharges vs. only the successful cases), the significant items rise to 25. Median p values under the two conditions are .31 and .15, in order.

Discussion

The differences in response distributions are small for examinees who were discharged before and after six months. Given that the significance of chi-square is inflated by large sample sizes, and that the probability of Type I errors is great in such a large set of significance tests, a finding of ten items out of sixty with $p < .05$ is not large. Also, given the small values of V for those ten items and the small between-group differences in response frequencies, the data do not support keying the instrument separately for the periods of initial and field service. As for causes of attrition, the very similar distributions of predictor responses for the two groups in this dataset do not imply that the reasons for early and late attrition differ.

Although the usefulness of keying long and short tenures differently is not supported, the value of using a longer criterion tenure for key development is. In the analyses here, almost twice as many items were sensitive to

real differences in success when the positive criterion group was purged of later attritions. The practice of developing keys on six month success seems here to undermine the validity of the predictor.

Conclusions

We now have evidence that traditional practices in developing biodata may have major flaws. A system for countering these problems is easy to conceive. Starting with a validated instrument, we could continually gather predictor scores of applicants and criterion scores of accessions. Today's selection measures would also be used as the predictor data for a later generation of scoring key, which would be based also on the performance measures. Updating of keys would then be ongoing rather than rare and ad hoc, as it is now. With ongoing updating, keys would be available after a minimal time lag and with appropriate generalizability (i.e., from applicants to applicants). Of course, increasing the criterion tenure would increase the time until new keys were available, but the best tradeoff between lag and quality could be determined empirically. Although problems in operating a biodata screen have been documented here, practical solutions are available.

References

Atwater, D. C. & Abrahams, N. M. (1983). Adaptability screening: Development and initial validation of the Recruiting Background Questionnaire (RBQ) (NPRDC TR 84-11). San Diego: Navy Personnel Research and Development Center.

Erwin, F. W. (1985). Development of new Military Applicant Profile (MAP) biographical questionnaires for use in predicting early Army attrition. Unpublished manuscript.

Frank, B. A. & Erwin, F. W. (1978). The prediction of early Army Attrition through the use of autobiographical information questionnaires (Technical Report No. TR-78-A11). Alexandria, VA: Army Research Institute.

Guilford, J. P. & Fruchter, B. (1973). Fundamental statistics in psychology and educaton. New York: McGraw-Hill.

Goodstat, B. E. & Yedlin, N. C. (1980). First tour attrition: implications for policy and research (Research Report 1246). Ft. Benjamin Harrison, IN: Army Research Institute.

Hicks, J. M. (1981, March). Trends in first-tour armed services enlisted attrition rates. Paper presented at the annual meetings of the Southeastern Psychological Association. Atlanta, GA.

Means, B. & Heisey, J. (1985, in press). Educational and biographical data as predictors of early attrition. Alexandria, VA: Human Resources Research Organization.

Walker, C. B. (1985). The Army's Military Applicant Profile: Its background and progress. In B. Means (Editor), Recent developments in military suitability research. (In preparation)

Use of Personal History Information to Predict Naval Academy Disenrollment

Joyce D. Mattson
Norman M. Abrahams
Rebecca D. Hetter

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Background

The Naval Academy has had great success in recent years in reducing its attrition from an average of about 36% for the Classes of '73 through '78 to the present rate of about 20%. Of the principal types of attrition -- academic and motivational -- motivational represents the largest single cause at 11%.

In an effort to reduce this motivational attrition further, the Naval Academy introduced an experimental background questionnaire initially developed in 1981. This instrument -- the Personal History Questionnaire (PHQ) -- is similar in its structured item format to instruments which have been useful in private industry for predicting such criteria as tenure, job performance, and creativity (Asher, 1972: Chaney and Owens, 1964: Schuh, 1971).

The questionnaire was administered to incoming Academy plebes in 1981 and 1982 (Classes of '85 and '86) and refined after each administration. Refinements were directed at: (1) improving item format, (2) retaining items which predicted attrition at the end of 1 year and (3) adding items suggested by the attrition analysis and a literature review (see Hetter and Neumann, in press). The final version of the questionnaire contains 85 multiple choice items, 25 of which were retained across all questionnaire revisions. The 85 items tap the factor-analytically-derived constructs in Table 1.

Table 1
Factor-analytically-derived PHQ Item Categories

Category	Number * of Items
Time management	6
Social comfort	6
Economic achievement orientation	7
Confidence in ability to handle military aspects of Academy	5
Degree of skill, participation in athletics	3
Persistence	6
Degree of financial self-support	4
Confidence in ability to handle academic aspects of Academy	3
Risk-taking propensity	5

Firmness of career goals	3
Degree of enjoyment of adventurous activities	3
Personal organization, self-discipline	4
Confidence that will manage life at the Academy well	3
Sociability	5
Belief in ability to control own destiny	7
Amount of parental education	2
Amount of reading individual does	2
Style of upbringing	5
Importance of being in a position of authority	1
Strength of accomplishment needs	1

* Several items were omitted from the analyses to prevent linear dependency.

Purpose

The purpose of this paper is to present the status of our research to date on the use of the Personal History Questionnaire to predict voluntary motivational attrition. In particular, this includes (1) development and cross-validation of an interim disenrollment scale using some of the plebe-based data collected for questionnaire development and (2) comparison of applicant and plebe administrations of the questionnaire to determine whether this plebe-based scale's validity would be likely to generalize to applicants.

Scale Construction and Validation

The interim disenrollment scale was developed using majority male plebes from the Class of '85 who had no prior military service and who completed the PHQ during their first week of plebe summer. This subgroup contains most Academy students but is homogeneous enough that important differences between attritees and non-attritees should be apparent.

Using end of second year attrition information, the item responses of voluntary motivational attritees (N=108) were contrasted with the responses of non-attritees (N=764). Sixteen response alternatives among the 25 items common to all PHQ revisions showed an 8% or greater endorsement difference and were scored on the scale.

The scale was then cross-validated on a dichotomous attrition criterion (voluntary motivational attrition vs. non-attrition) using individuals from the Classes of '86 and '87, who where also administered the PHQ as plebes. Both first- and second-year attrition information was available for the Class of '86, but only first year attrition information for the Class of '87. All racial, gender, and prior-service-experience subgroups were combined, since operational predictions would be required for all types of applicants.

Table 2 shows the results of these cross-validity analyses --- the biserial correlations of the disenrollment scale with cumulative voluntary motivational attrition at different tenure points. Biserial correlations were computed under the assumption that the attrition dichotomy actually reflects an underlying normally distributed continuum of motivation (McNemar, 1969).

Table 2
Relationship of the Disenrollment Scale
to Cumulative Voluntary Motivational Attrition

Academy Class	Attrition status as of the end of:	Sample Sizes: Total	Non-attritees	VM* Attritees	r bis**
'86	Plebe summer	1158	1125	33	.35
	Year 1 Semester 1	1144	1071	73	.26
	Year 1 Semester 2	1113	1032	81	.28
	Second summer	1112	1028	84	.26
	Year 2 Semester 1	1100	1004	96	.27
	Year 2 Semester 2	1078	958	120	.26
'87	Plebe summer	1341	1281	60	.33
	Year 1 Semester 1	1315	1208	107	.32
	Year 1 Semester 2	1291	1166	126	.26

*VM = Voluntary motivational.
**Computed by contrasting VM attritees with non-attritees at each successive tenure point.

The correlations in both the '86 and '87 classes are in the .26 to .35 range, indicating reasonably good differentiation between motivational attritees and non-attritees. While correlations computed on the Class of '86 may be inflated by the use of '86 first-year attrition information in revising the PHQ, correlations computed on the Class of '87, which have no such contaminant, are remarkably similar.

Assuming for a moment that validities derived on plebes could be generalized to the selection situation, Table 3 suggests the level of improvement in the prediction of voluntary motivational attrition which could result from adding the PHQ scale to the currently used Candidate Multiple.

Table 3
Incremental Validity of the Disenrollment Scale
for Predicting Voluntary Motivational Attrition

Year Group	Attrition status as of the end of:	N	Candidate Multiple r bis *	Candidate Multiple Plus Disenrollment Scale R bis *
'86	First year	1109	.06	.28
	Second year	1076	.12	.28
'87	First year	1288	.00	.26

Based on these data, the validity against first- or second-year voluntary motivational attrition could be increased substantially by the scale's use. Two caveats are in order, however. First, these increases represent upper limits, since in practice the Naval Academy would be unlikely to weight the disenrollment scale as heavily as is reflected here. Second, even the maximum gain would be slightly less if validities were corrected for range restriction.

Of interest, also, from an operational perspective, is the relationship of the disenrollment scale to other criteria which the Academy attempts to predict -- namely, military quality point ratio (MQPR), academic quality point ratio (AQPR), and academic attrition.

Table 4 shows that the scale has a slight positive relationship to MQPR, and no relationship to AQPR, suggesting that no deterioration in these performance areas would be likely to result from the scale's use. The negative relationship which is apparent with first-year academic attrition becomes slightly positive over the longer perspective of two years, suggesting that the long-term academic attrition rate would not be adversely affected by the scale's use, but that some of the attrition might occur earlier than at present.

Table 4
Relationship of the Disenrollment Scale
to Academic Attrition, AQPR, and MQPR

Criterion	Academy Class	Attrition status as of the end of:	N	* r
Military QPR	'86	Year 1 Semester 1	1092	.05
		Year 1 Semester 2	1066	.08
		Year 2 Semester 1	1021	.08
		Year 2 Semester 2	992	.08
	'87	Year 1 Semester 1	1235	.04
		Year 1 Semester 2	1187	.07
Academic QPR	'86	Year 1 Semester 1	1092	.01
		Year 1 Semester 2	1066	.03
		Year 2 Semester 1	1021	.01
		Year 2 Semester 2	992	.01
	'87	Year 1 Semester 1	1237	.01
		Year 1 Semester 2	1187	.01
Academic Attrition	'86	Year 1 Semester 1	1083	-.20
		Year 1 Semester 2	1064	-.00
		Year 2 Semester 1	1043	.03
		Year 2 Semester 2	1017	.05
	'87	Year 1 Semester 1	1226	-.20
		Year 1 Semester 2	1205	-.15

*Biserial correlations are used for the dichotomous attrition criterion; Pearson product-moment correlations for the other continuous criteria.

Generalizability of Validities to Applicants

To assess the likelihood that the disenrollment scale's promising validities would generalize to applicants, 1111 members of the Class of '88 who completed the PHQ as applicants were retested as plebes. If the two occasions yielded similar disenrollment scale scores, a plebe-based scale would probably retain its validity when applied to applicants; if the two occasions yielded markedly different responses, the scale's validity for selection would be uncertain.

Table 5 shows mean scores and applicant- plebe correlations on the disenrollment scale for the Class of '88.

Table 5
Means, S.D.s, and Applicant-plebe Correlation for the Class of '88 on the PHQ Disenrollment Scale (N=1111)

Sample	Mean	S.D.	r
Applicants	4.875	2.566	.52
Plebes	3.319	2.892	

The applicant means are more than 1/2 standard deviation higher than the plebe means, suggesting that applicants may slant their responses to increase their chances of Academy selection. At the item level, this distortion is apparent in the following representative item:

Item and Alternatives	Percent Selecting Alternative: Applicants	Plebes
I find myself putting things off until the last minute:		
A. Almost always	0	3
B. Often	10	25
C. Sometimes, but not often	48	50
D. Rarely	37	20
E. Never	5	3

In addition to item response distortion, the rank-ordering of individuals on the disenrollment scale also changes, as reflected in an applicant-plebe correlation of only .52. Successively removing items with the lowest applicant-plebe correlations from the scale improves this correlation by only .03.

These results indicate that individuals answer the PHQ differently as applicants than they do later as plebes and that validities based on plebes cannot necessarily be assumed to generalize to a selection situation. Rather,

plebe-derived scales must be validated on applicants or new applicant-based scales must be constructed.

Summary

In summary, results of this research suggest that:

(1) The items of the Personal History Questionnaire, when completed under low incentive to distort, car distinguish between individuals who leave the Academy for voluntary motivational reasons and those who remain.

(2) Individuals complete the PHQ differently as applicants than they do as plebes, in a way suggesting a tendency to slant applicant answers to increase the chances of Academy selection.

(3) The amount of change varies for different individuals so that applicant and plebe administrations yield different rank-orderings on the disenrollment scale and validity results from plebes cannot be assumed to apply to the selection situation.

REFERENCES

Asher, J.L. The biographic item: can be improved? Personnel Psychology, 1972, 25, 251-269.

Chaney, F.B. and Owens, W.A. Life history antecedents of sales, research, and general interest. Journal of Applied Psychology, 1964, 48, 101-105.

Hetter, R. and Neumann, I. (in press). Preliminary development of a Personal History Questionnaire for predicting Naval Academy voluntary disenrollment. (NPRDC in press). San Diego: Navy Personnel Research and Development Center.

McNemar, Q. Psychological Statistics. New York: John Wiley & Sons, 1969. 4th Ed.

Schuh, A.J. The Predictability of employee tenure: A review of the literature. Personnel Psychology, 1967, 20, 133-152.

Intercorrelations of Biographical Information, Aptitude Test Scores and Job Performance Ratings for 108 Occupations

Marvin H. Trattner
Office of Personnel Management, Washington, D.C.

In this paper General Aptitude Test Battery (GATB) scores, age, amount of both occupational experience and plant experience, employee sex, education level and job performance ratings are intercorrelated for employees in 108 occupations.

The research was accomplished to serve the following purposes:

1. To calculate and summarize a large number of validity coefficients for biographical variable and aptitude test predictors of job success. Validity coefficients for biographical variables are generally not as plentiful in the general literature as for paper and pencil tests. The Schmidt-Hunter validity generalization procedure is applied to the validity coefficients in order to summarize them and correct them for known artifacts such as criterion unreliability and predictor restriction-in-range.
2. To calculate the unique contribution to valid prediction of the biographical variables.
3. To determine whether predictor validity is related to job complexity level.

Background

USES GATB testing and validation program. The United States Employment Service (USES) developed the GATB approximately 45 years ago for selecting and counseling applicants for vacancies filled by state employment services. The state employment services generally fill vacancies in low level industrial and clerical occupations. The GATB employs 12 subtests to measure the following nine relatively independent abilities (Anastasi, 1982):

G - Intelligence	Q - Clerical Perception
V - Verbal Aptitude	K - Motor Coordination
N - Numerical Aptitude	F - Finger Dexterity
S - Spatial Aptitude	M - Manual Dexterity
P - Form Perception	

Two and one-half hours are required to complete the GATB. The GATB subtests were developed on the basis of factor analysis, the aim was to employ subtests that intercorrelated only minimally except for the Intelligence aptitude which is measured by the sum of three subtests also used in part to measure Verbal, Numerical and Spatial aptitudes. The GATB tests applicants for most occupations in the U.S. economy hence it measures a very wide range of ability. It is similar in this respect to the Armed Forces Qualification Test.

The opinions expressed in this paper are those of the author, who is solely responsible for the accuracy of the contents, and does not necessarily reflect the views of the Office of Personnel Management.

The USES has been conducting validity studies on the GATB since its development. The measure of job performance which served as the criterion in these 108 studies were the summed ratings of job performance made by two supervisors. All studies used the same standardized performance rating form. Supervisors rated employees on 6 five-point scales. The scales were used to evaluate employees on the quantity, quality and accuracy of their work, their job knowledge, the variety of duties performed and also on overall job performance.

The data reported here were collected since the early seventies. Subjects were cumulated across geographical locations by occupation. The average number of cases per occupation was 222, a relatively large number when compared to the typical n employed in civilian validity studies. A concurrent validity design was used. That is, current employees in the occupation were used as subjects. The correlations obtained were assumed to apply also to job applicants except there is known to be a restriction-in-range in predictor scores which lowers obtained correlations and requires the use of a correction factor.

Schmidt-Hunter Validity Generalization Procedure (VGP). Schmidt and Hunter (1977) developed a procedure for summarizing validity coefficients for specific predictor/occupation categories. Previously no statistical procedure was available to summarize and correct a group of validity coefficients for known artifacts (such as sampling error, predictor restriction-in-range, and criterion unreliability) in order to estimate both the mean true validity and the standard deviation of the true validity coefficients. They and their associates employed the VGP to demonstrate in a large number of recent studies that predictor validities are generally much more consistent and of higher magnitude than was previously thought.

The VGP is accomplished by performing the following calculations on uncorrected validity coefficients obtained with a concurrent validity strategy.

1. Validity coefficients for a specific predictor/occupation combination are collected.

2. The mean true validity for the predictor/occupation is then calculated by first calculating the n weighted mean of the obtained validity coefficients. This mean is then corrected for attenuation for mean estimated criterion unreliability and also for mean estimated restriction-in-range in the predictor. If no criterion reliability or predictor restriction-in-range data are available for the validity studies at hand then hypothesized distributions are employed.

3. The standard deviation of the true validity coefficients is calculated next. First, the n weighted standard deviation of the obtained validity coefficients is calculated. Then the following extraneous sources of variance are subtracted from the variance of the obtained validity coefficients, the variance due to a) average sampling error b) differences in studies due to specific predictor reliability, criterion unreliability, and

restriction-in-range in the predictor. The remaining variance is then corrected upward by the same correction factor that was used to correct upward the mean obtained validity in order to estimate the mean true validity.

Additional considerations. A modification was required in the VGP for this study because the same specific predictor and the same criterion rating form was employed in every validity study in the data set. It was decided since the only predictor for any predictor/occupation combination was the specific GATB subtest that no predictor unreliability variance would be removed from the obtained validity variance. The criterion rating form was only one of many sources of criterion unreliability, consequently variance across studies in validity due to differences in criterion unreliability was removed from the obtained validity variance. For all predictors the standard hypothesized predictor restriction-in-range distribution was employed to correct the obtained coefficients. The mean selection ratio for this distribution is .45. In the latter two cases, the standard validity generalization procedure was followed.

The occupations included in the study carried Dictionary of Occupational Titles (DOT) codes. The DOT codes contain no job complexity designation. A job complexity score was needed since it has been frequently found that job complexity moderates predictor validity. A job complexity score was obtained for the 108 occupations by summing scaled scores for estimates of required mathematical development, language development and specific vocational preparation (U.S. Department of Labor, 1981). The complexity score sums for the 108 occupations ranged from scores of 4 to 16. Most of the occupations for which data were available were low level clerical and industrial jobs. The two occupations receiving the highest complexity score were Medical Technologist and Electronics Technician. Some examples of occupations receiving the lowest score were Hard Packager, Garment Folder, and Small Product Assembler. The distribution of complexity level was positively skewed, most of the occupations clustered at the lower complexity levels.

RESULTS AND DISCUSSION

There were individual score data for 23,917 employees in 108 occupations. Complete data were available for all subjects.

Mean true validity for each predictor for all occupations combined. All predictors (GATB test scores and biographical variables) and the criterion score were intercorrelated separately for each occupation. The obtained validity coefficients were then combined for all occupations for each predictor. The mean and the standard deviation of the distribution of obtained validity coefficients was calculated for such predictor. The VGP was then applied to the distribution of 108 obtained validity coefficients for each predictor to calculate the mean true validity, the standard deviation of the true validity coefficients and the "95% credibility value" (the point above which 95% of the true validity coefficients lie). Table 1 contains the obtained and true validity coefficient distribution statistics for each predictor.

Table 1

Mean and Standard Deviation of the Obtained and True Validity Distribution for the Predictors

	Predictors													
Obtained Validity Distribution	Sex[a]	Age	Educat. Level	Plant Exper.	Occup. Exper.	GATB Aptitudes G	V	N	S	P	Q	K	F	M
Mean	-00	07	02	19	19	20	16	20	13	14	16	08	09	10
Standard Deviation	08	11	10	11	10	09	08	08	08	08	08	08	08	09
True Validity Distribution														
Mean	-01	16	05	39	39	43	34	42	28	29	33	18	20	22
Standard Deviation	10	19	15	15	15	06	09	04	10	09	07	12	09	12
95% Credibility Value	-18	-16	-20	14	15	32	21	35	12	14	22	-01	04	02

note: Decimal point omitted for validity coefficients.
a: Scored 1 = male, 2 = female.

Table 1 reveals that when all occupations are grouped together all tests have higher mean validity than the mean validity for sex, age or education level. Both plant experience and occupational experience correlate relatively highly with job performance ratings, only intelligence and numerical ability have higher mean true validity. All the biographical predictors have considerably higher true validity variance than the test predictors.

When the mean predictor intercorrelations were clustered across the 108 occupations the following four groups of predictors were obtained - experience cluster (age, plant experience, occupational experience), cognitive ability (intelligence, verbal, numerical), perceptual ability (spatial, form perception, clerical perception), psychomotor ability (motor coordination, finger dexterity, manual dexterity).

Unique contribution to predictor validity of the experience cluster) for three job complexity categories. The unique contribution to validity of the summed standard scores for age, plant experience and occupational experience when added to the cognitive ability score (sum of G, V, N) was determined separately for high, middle, and low complexity occupations. The calculations were made by first averaging the correlations for the occupations separately for the three complexity categories. The thirteen job complexity categories were compressed to three by combining complexity levels 4-8, 9-11, and 12-16. Employee sex and educational level were eliminated from consideration since they correlated so poorly with the criterion. All variables were unit weighted and standard scores were employed. The calculations were made for the obtained validity coefficients rather than the true validity coefficients. Cognitive ability correlated highest with the criterion and was selected first, except for the low complexity level where it equalled experience.

Table 2

Obtained Correlations of Sum of Unit Weighted Standard Score Cognitive Ability and Experience Predictors for Three Job Complexity levels.

Job Complexity Level	Predictor Composite Employed G, V, N, Age Plant Exp. Occupa. Exp.	G, V, N	Unique Validity of Experience Cluster
Low	.271	.180	.091
Medium	.294	.245	.049
High	.270	.195	.075

Table 2 reveals that the experience cluster adds a considerable amount of unique validity to the validity obtained by the cognitive predictors for all three job complexity levels. The unique validity of the experience cluster is identical to the part correlation of the experience cluster with the criterion. Neither the perceptual or psychomotor ability clusters add unique validity to the cognitive ability cluster. No great differences in validity occurred at the three complexity levels.

Consistency of validity coefficients across job complexity levels. The mean true validity for each predictor was calculated separately for each job complexity level. The product moment correlation of the mean true validity coefficients with the job complexity scale for each predictor confirmed some foreseeable relationships. The following predictors had higher validity in lower complexity occupations: perceptual and psychomotor abilities, age, plant experience, occupational experience. Education level had higher validity in higher level occupations. Women were rated higher than men for lower level occupations and men were rated higher than women for higher level occupations. However, it is important to emphasize that while many of these correlations between validity and complexity level were significant, in general, they would not lead to the use of different predictors for different complexity levels. For instance, although education level is correlated higher with job performance for more complex jobs, the validity for the more complex jobs is not high enough to add unique variance to the cognitive ability score.

I have reported results that showed a lack of important moderation of validity by job complexity in two different contexts. Approximately the same validity was obtained for the predictor clusters for the three different complexity levels and validities for the specific complexity levels show that very rarely would a different predictor be used than the ones that applied to all complexity levels combined. This may be because these data were cumulated over geographical locations which is known to depress the validity coefficients.

Summary

The following results were obtained:

1. GATB tests in general had higher correlations than biographical variables with supervisory ratings. Two biographical variables, plant experience and occupational experience however had high correlations with supervisory ratings.

2. The experience cluster score, consisting of plant experience, occupational experience and age added unique valid variance to the cognitive ability cluster score. Neither perceptual nor psychomotor ability added valid variance to cognitive ability.

3. The data show some relation between predictor validity and job complexity level but not enough to justify the use of different predictors for different complexity levels. A reason for this lack of relationship was advanced.

REFERENCES

Anastasi, A. (1982). Psychological Testing. New York: Macmillan, 5th edition.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

U. S. Department of Labor Employment and Training Administration (1981). Selected characteristics of occupations defined in the Dictionary of Occupational Titles. Washington: U. S. Government Printing Office.

ANALYTIC PREDICTION OF TRAINING DEVICE EFFECTIVENESS

Louise G. Yates
Douglas Macpherson

U. S. Army Research Institute

One goal of the U. S. Army is to design cost effective training equipment which prepares troops to do necessary tasks. The design and acquisition of this equipment is usually accomplished during the design and acquisition phases of the parent weapon system. Although the training equipment design is supposed to be based on field tests, often the field tests are eliminated due to logistical problems related to the parent equipment acquisition and design process. When field tests are performed, they often come too late in the acquisition process, due to the same logistical problems, to allow change of the device design based on the results of the field test. Thus, device design often becomes dependent on the analytic procedures of each program manager as no formal analytic procedures for evaluation of training device design have been implemented.

The Army Research Institute has designed the Device Effectiveness Forecasting Procedure (DEFT) to fill the need for a formal analytic procedure that can be used at various stages in the design/acquisition process to evaluate the proposed device design, proposed design changes, and alternate device designs. The instruction and training principles contained in DEFT can also be used as a guide in the initial design process. DEFT is based on a review of relevant literature and on the development of a conceptual approach which takes into consideration theoretical and practical issues of training device design, development, and evaluation. This approach led to a program development model which resulted in a network of hypotheses that relate program inputs to intermediate outcomes.

DEFT is computerized in an interactive menu-driven format for use on the IBM PC or compatible computers. It requires ratings from subject matter experts (SME) on device, operational equipment, trainee, and training system characteristics. It can be used at three levels of complexity , DEFT I, II, and III, varying with the amount of information available to the user. DEFT I is designed for use early in the design process when only general information is available to the user. At this early stage alternate device designs may need to be compared or one device design evaluated. An evaluation made at an early stage is especially valuable as design changes have much less cost and time impact.

DEFT I requires eight global ratings from the SME based on general information about characteristics of the device, the parent equipment, the training needed, the trainees, instructional features of the training, and the physical and functional similarity of the device to the operational equipment.

DEFT II and III are designed to be used later in the design/acquisition cycle when more detailed information is available. This can be the first use of DEFT, or the use of DEFT II or III to check the forecast made by DEFT I , or the use of DEFT II or III to determine the effect of proposed changes in the device. DEFT II ratings are made at the task level. There are thirteen kinds of ratings, some of which are rated for each task trained on the device or the operational equipment.

DEFT III can be used at either the task or subtask level. It requires 35 different types of ratings, some of which are required for each task/subtask level. Added to the information needed to make DEFT I and II ratings are certain features of controls and displays on both the device and the operational equipment.

DEFT II and III provide more diagnostic capability than does DEFT I. With DEFT II and III it is possible to predict the impact of changing one or more specific characteristics of the device or the training system on the effectiveness of the training. DEFT III which requires the most information can provide more diagnostic information than DEFT II.

The information input into DEFT is converted into several numerical indices which are used to estimate device effectiveness. Table I contains the evaluation summary from a DEFT I analysis of a naval training simulator for the SH-3 helicopter.

The indices listed in the summary are the same for DEFT II and III. The numbers in the left column come directly from the ratings (For DEFT II and III these numbers are the ratings summed and averaged over tasks/subtasks and rating scales.). Indices in the other two columns are derived from mathematical combinations of the numbers to their left and above them in the table. The three summary indices (right column) are for acquisition effectiveness, transfer effectiveness and the sum of these two, total effectiveness.

Training Problem defines the deficiency in skills and knowledge that the new trainees have relative to criterion performance on the training device and the difficulty trainees would have in overcoming the deficiency. Training Efficiency is rated on the basis of the instructional features and training principles incorporated in the training device. Acquisition Effectiveness is designed to give a "poor" score to a device that has a large training problem and is inefficient in dealing with it.

TABLE I

DEFT I Evaluation Summary

Performance Deficit	50		
Learning Difficulty	60		
Training Problem		30.00	
Quality of Training Acquisition	90		
Acquisition Efficiency		.94	
Acquisition Effectiveness			31.91
Residual Deficit	20		
Residual Learning Difficulty	30		
Physical Similarity	80		
Functional Similarity	80		
Transfer Problem		6.00	
Quality of Training Transfer	80		
Transfer Efficiency		.89	
Transfer Effectiveness			6.74
Total Effectiveness			38.65

The Transfer Problem is defined by the amount of learning required on the operational equipment, or by other means, after training on the device is completed, ratings of the difficulty of this learning, and ratings of the functional and physical similarity of the device controls and displays to those on the operational equipment. The Transfer Efficiency Index is computed from ratings of the instructional and training principles in the device that contribute to transfer of training.

Like the Acquisition Effectiveness Index, the Transfer Effectiveness Index gives a "poor" score to a device that has a large transfer problem and is ineffective in dealing with it.

Evaluation of DEFT

At the time that DEFT was completed a field evaluation could not be funded. Therefore, an analytic assessment was made to assess certain scalar properties of DEFT and to examine inter-rater agreement. Based on expected score distributions obtained from Monte Carlo distributions, devices with different DEFT ratings can be evaluated as different or not. Based on certain reasonable assumptions regarding the distribution of expected input values, DEFT outputs are interpretable and meaningful. Sensitivity analysis, which tested the impact on DEFT output of

varying input values, indicated that, with the exception of the two efficiency scales, all input values had the same effect on total score. Efficiency scales had a larger effect than other scales.

Interrater agreement was examined by having four raters use DEFT I, II, and III to evaluate three training devices, the M-60 Gunnery Trainer (VIGS), the Burst-on-Target (BOT) Trainer, and the Maintenance Procedure Simulator (MPS) for the E-3A Navigation Computer System. The seven summary indices shown in the two columns on the right of Table I were computed for each DEFT rater and trainer combination. From an analytic assessment of these results it was concluded that there was substantial interrater agreement for all DEFT indices across the three devices. However, it is necessary for raters to agree on their assumptions regarding the device, trainee population , device utilization, and the meanings of the various DEFT scales prior to making the ratings.

Funds have been allocated for an interservice contract effort, starting in FY86, to field test DEFT. SME ratings of a number of training devices with known performance capabilities will be used. One output expected from this effort is the conversion of pertinent DEFT index scales to scales based on meaningful training concepts such as trials to mastery, training time, etc.

Wide interservice interest has been displayed in DEFT. Personnel from the Training Analysis and Evaluation Group (TAEG) of the Naval Training Evaluation Center have used DEFT I and II to evaluate the 2F64C helicopter trainer, under four different conditions, to determine if DEFT in its present form will be useful to TAEG. Interrater correlations for two raters using DEFT II ranged from .81 to .97 for the seven groups of ratings which result in the seven summary indices of Table I. Correlation of DEFT I and II measures of transfer efficiency with transfer ratios based on trials to mastery in the actual helicopter, after training on the simulator, averaged .54. The Ns were small for all of these results, so the results are not conclusive, but they are encouraging.

In the near future the Marine Corps will be evaluating DEFT for their use by using it to evaluate the LVT Landing Vehicle Trainer. In addition, the Air Force has requested information on DEFT.

The interest of these agencies in DEFT appears to confirm the need for such a technique, and initial results of the evaluation of DEFT, although incomplete, lead us to believe that DEFT can become a very useful tool in designing and evaluating training devices.

Bibliography

Rose, A. M., Wheaton, G. R., & Yates, L. G. Forecasting Device Effectiveness: Volume I. Issues. Technical Report 680. Alexandria, VA: U. S. Army Research Institute.

Rose, A. M., Wheaton, G. R., & Yates, L. G. Forecasting Device Effectiveness: Volume II. Procedures. Research Product 85-25. Alexandria, VA: U. S. Army Research Institute.

Rose, A. M., Wheaton, G. R., & Yates, L. G. Forecasting Device Effectiveness: Volume III. Analytic Assessment of Device Effectiveness Forecasting Technique. Technical Report 681. Alexandria, VA: U. S. Army Research Institute.

BEST COPY AVAILABLE

Research on Decision Aids for Training Design and Evaluation

Angelo Mirabella
Army Research Institute
for the
Behavioral and Social Sciences

A key mission of the Army Research Institute (ARI) is to produce decision aids which will help training developers design and evaluate media, methods, and programs of instruction. This paper summarizes a program of research (focusing on maintenance training) to achieve that mission. The summary will address some research products which have been completed, and some which are still being developed. It will describe cooperative efforts among the services.

The programs long-range objective is to combine a number of job aids into a decision-support system (DSS) which will unburden the training developer. The DSS should provide an audit trail for his or her decisions, instill confidence that cost-effective training is being developed, improve communication among developers, and yet be easy to understand and use. Yet, products which are easy to use often require very complicated science and technology to develop as well as substantial investments in time and money. Therefore, we have set an ambitious, high risk goal. Why should we pursue it and what obstacles do we face? First, why do we need job aided, perhaps high technology, approaches to training development?

We need them to support increasing requests for accountability from the Congress. Congress has asked the Services to do a better job of planning, explaining, and defending how they train their respective forces. OSD and the Services have responded with new policies. For example, recent directives from OSD on training device acquisition and computer-based instruction, require that the choice of a training medium, its specific design and use in programs of instruction be justified. This requires data on the expected and/or measured effectiveness of the medium. At the very least, training managers and weapon developers will have to point, increasingly, to authoritative bases for decisions on training system design.

At the grass roots level, we need decision aids to support the Service schools and their training directorates. The schools face a very difficult task in developing training not only for current weapon systems, but for those on the drawing board. Their difficulty is exacerbated by shortages of training development specialists and turnover among those that are available. One result is that the schools develop training partly on the basis of intuitive judgement, which may be difficult to explain or justify. Therefore, the need for decision aids, if not a decision support system, is evident. Why haven't we met this need before? The answer lies in chronic gaps in the science and technology of training design, particularly for maintenance.

1. Lack of a critical mass of scientific and expert data to support training design decisions, and

2. Lack of methodology for translating scientific and expert data into design technology. These will continue to be chronic until a concerted and systematic effort is made to deal with them. Working towards a decision support system forces the issue for reasons to be explained shortly.

The Program

Our program at ARI is designed to help move the Army from a paper-based, approach to developing training to a higher technology, more objective, more easily documented, and more defensible approach (Figure 1). We have a systematic, institutionalized approach in Instructional Systems Design (ISD). But ISD has not been very successful, because it places a heavy information overload on the training developer and requires a many subjective judgements that are difficult to make. A decision support system could shift much of this burden from human to machine.

JOB AIDS FOR TRAINING SYSTEM DESIGN, DEVELOPMENT, ACQUISITION

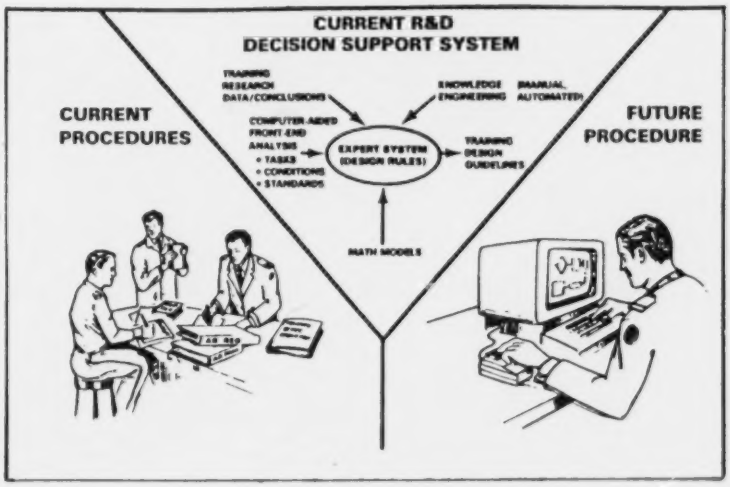


Figure 1

The middle of Fig. 1 illustrates how we can bring about this kind of magic. This section shows one kind of DSS, based upon artificial intelligence in the form of an expert system. Here, the training developer would communicate to the computer the maintenance tasks to be trained, conditions of training, and standards of achievement, and receive a set of recommendations for designing a maintenance simulator. The designer would have to provide the front-end analysis, i.e., tasks, conditions, and standards. But once this was done the design decisions would be made automatically by a set of built-in "if..then rules." Moreover, job aids can and have been developed to assist with front end analysis, so even this human function could be machine assisted.

Having decided to bring about this kind of magic, what implications does the decision have for R&D strategy? The answer is provided by outlining the steps required to construct a DSS. These provide the skeleton for the research program which I am addressing in this paper.

Step One is to accumulate a relevant set of scientific and expert data. In our case, that means data which show how various training variables influence the acquisition and transfer of maintenance skills. It also means some form of accumulated wisdom from training experts with much experience. We are in trouble on both counts, and we're only on Step One. For many reasons, we lack adequate amounts and kinds of the first set. But we're also in trouble on the second set because the training business does not possess the kinds of expertise that one finds in more mature disciplines like engineering, or medicine.

Step Two is to organize the scientific data so that it is easy access and process. The traditional approach has been to store scientific data in journals which are often difficult to obtain and understand. Data Base Management System technology offers an alternative approach.

Step Three is to develop and assess a variety of technologies for processing the data in Step Two so that we can generate useful conclusions and principles of training design. Many technologies are available. But while they work well in physical science, engineering, and medicine, they have not yet been exploited for training problems. Examples of these technologies include the methods of operations research, mathematical modeling, pattern recognition techniques, decision analysis, and artificial intelligence. In Step Four, the results of the first three steps have to be packaged for easy use by the ultimate customer.

At ARI we have been or plan to do some work under each of the four headings above. For example, under Step One we have set up a maintenance training research facility at George Mason University which is somewhat unique. We're using Army simulators and studying Army maintenance tasks (diesel and Hawk radar repair), but under the controlled conditions of a laboratory. Over the next three years, we hope to accumulate, a systematic collection of training effectiveness data which can be converted into

specifications for designing simulation and computer-based instruction of maintenance skills. These in turn would provide part of the basis for an expert system we are developing.

We have completed the first two of approximately 10 major experiments. We will be looking at the effects of alternative media, (e.g. computer-aided instruction vs simulator training), training strategies, and individual differences. We want to identify the most potent training system variables and translate them into design principles. Initial results have been described by Swezey (1985). Under Step Two we have constructed a preliminary training simulation research data base. It contains of 150 research papers which have been analyzed into 20 categories of information. The data exists in hard copy (Ayres, et al, 1984) and exists in computerized form. We plan to conduct a meta-analysis on an expanded and revised version of the data base and use the results as inputs to our expert system. This brings us to Step Three: developing techniques for converting data into useful decisions.

As part of Step Three we have developed procedures to help design and evaluation of training. The evaluation aids are finished or nearly finished products ready to implement, so I will describe them first. One of these products, a method called Comparison-Based Prediction of Training Device Effectiveness, is based upon Comparability Analysis (CA) which is currently in use by the military -particularly the Air Force - for estimating the reliability of new weapon systems. We have adapted CA for use in training effectiveness analysis. The analyst uses his or her knowledge (or performance data if available) about existing training devices to estimate the effectiveness of new devices. A handbook, explaining how to conduct the analysis has recently been produced (Klein et al, 1985) and we are currently exploring ways to field test and implement it.

We have also developed an alternative - and in this case computerized - evaluation procedure called the DEFT - Device Evaluation Forecasting Technique (Rose et al, 1985a). DEFT is a decision analytic method - as is CBP, but one which is based on a mathematical model. The analyst is led through questions about the device being evaluated. His or her answers (judgements on a scale of 0 to 100). provide the inputs for formulas which yield various measures of device effectiveness. These, in turn, are combined mathematically into an measure of device effectiveness. The method is being used on a trial basis by the Naval Training Equipment Center and will be extensively demonstrated and field tested during the current fiscal year under a new program called ASTAR (Automated Simulator Test and Assessment).

Lastly, we have developed a draft guidebook on how to empirically test simulators and training devices for transfer of training effectiveness. This document is designed to help action officers who are responsible for evaluating training devices but who may not be experts in testing. It aims to teach the reader enough about the basic concepts, language and methods of testing so that he or she can communicate with

evaluation specialists and assess their recommendations. It is also designed to help that action officer argue for adequate testing resources. We have made preliminary plans to hold a workshop for the Army Training Device Community.

Currently, our Step Three work is devoted to building expert system for design of training simulators. So far we have built a demonstration system containing 60 rules. We'd like to triple the size of the demonstration and then take it to a TRADOC School to develop it further with the help of maintenance instructors.

Recently we began a joint program with NTEC and AFHRL to develop automated techniques for producing training courseware. Our objective is to construct models of students, instructors, and subject matter experts and then convert these using artificial intelligence methodology into automated procedures for doing training front end analysis and building diagnostic routines. We plan to test the results of this re search on a training problem that holds common interest for the three services. We anticipate combining training evaluation methods, the expert system for training device design, and the automated procedures for courseware development into a comprehensive decision support system for training design.

Under Step Four we have written some concept papers identifying and defining issues in the human engineering and design of a decision support system (Macpherson & Mirabella, 1985; Singer & Hays, 1985). We are also developing a joint program with the Ordnance School so that we can conduct research on user requirements and test approaches to meeting those requirements, including specific human engineering designs for the decision aiding technology that evolves from our various lines of research.

Future Plans and Research Pay-offs

Where do we go from here? We plan to make a concerted effort to field test, demonstrate, and implement the training gvaluation tools already completed. We will also continue research on artificial intelligence approaches to training development (expert system and automated courseware procedures). Finally, we will be developing a test site where we can study user requirements and match them with the technologies evolving from our basic research program. We are targeting FY87 to complete the system components, FY88 to integrate them, and FY89 to demonstrate the resulting DSS at the test bed site.

Suppose we succeeded beyond our wildest dreams and did, in fact, evolve a DSS which could significantly unburden the training developer. What benefits would accrue? How could we use such a product to help the Army meet its future maintenance training rmquirements? First, it would contribute significantly towards meeting Congressional and OSD/Service directives for accountability in developing, executing, and managing training. Secondly, it could help document future training requirements.

A projection of future population characteristics could be used as input data for the DSS, and if it were sufficiently comprehensive, it could return specifications for appropriate future training system designs. If, on the other hand, we discovered that a particular future projection could not be represented in the DSS, we could then document this state of ignorance and use that documentation as a guideline to new research. With such a system, we should also be able to significantly reduce the problem of orienting future action officers who will be faced with the responsibility for training development. Finally, we could cure the chronic ailment of shortages in training and educational specialists.

Whatever the future holds in store because of changes in Army missions, technology, resources, or population characteristics we will be better able to meet that future if we close critical gaps in the science and technology of training design. The concept of the DSS provides a useful way to organize the research and development which will be required to close those gaps.

References

Ayres, A., Hays, R.T., Singer, M.J., & Heinicke, M. An Annotated bibliography on the use of simulators in Technical Training. ARI Research Product 84-21, October, 1984.

Klein, G.A., John, P.G., Perez, R., & Mirabella, A. Comparison-Based Prediction of Cost and Effectiveness of Training Devices - A Guidebook. ARI Research Product, In press, 1985.

Macpherson, D.G. & Mirabella, A. APPLIED META-ANALYSIS: A Procedure for Speeding Innovation by Transferring Scientific Knowledge More Quickly. In "Artificial Intelligence and Simulation", William M. Holmses, Ed. San Diego: Simulation Councils Press, 1985a.

Rose, A.M., Wheaton, G.R., & Yates, L. Forecasting Device Effectiveness: Procedures. ARI Research Product 85-25, June, 1985.

Singer, M., and Hays, R. New Directions in Simulation Research: Generating, Handling, and Delivering Empirically Based Guidance. ARI Technical Report RR 1389, January, 1985

Swezey, R.W., Criswell, E.L., Huggins, R.S., Hays, R.T., & Allen, J.A. Training Effects of Hands on Practice and Three Instructional Methods on a Simple Procedural Task. ARI Technical Report, 1985.

THE ARMY EXPERIENCE SURVEY: METHODOLOGICAL HIGHLIGHTS

Jeanna F. Celeste
Westat, Inc.
1650 Research Boulevard
Rockville, Maryland 20850

Forecasting future enlistment and reenlistment trends, and developing effective recruitment and reenlistment policies and procedures has become a very complex task for the all-volunteer forces. Recent upturns in the state of the U.S. economy have placed military recruiters in competition with one another and with the civilian workplace for the limited labor resources. Military policy makers regularly need current information about the characteristics of people enlisting as well as those separating in order to make informed decisions. More importantly, they need to understand the reasons for enlistment and separation.

The Army Research Institute for the Behavioral and Social Sciences,* recently awarded a contract to Westat, Inc. to design and administer a large random survey of Army veterans. A major impetus for performing the survey was the interest expressed by the Secretary of the Army in assessing the post-service experiences, plans for reenlistment or reserve duty, and attitudes toward military experience of recently separated soldiers.

It is hoped that the responses of recently separated Army veterans can provide important data to illuminate their expectations and values concerning civilian life and to detail gaps between their expectations versus experiences in their recent period of active duty. In addition, when compared to data from Army subgroups who elected to reenlist, these data can illuminate differences in the expectations/values across various population subgroups.

This paper reviews the approach taken to obtain these data. Sampling and survey administration procedures are described along with a brief discussion of survey outcomes.

Survey Administration Procedures

The Army Experience Survey was comprised by a broad variety of tasks. A sample of recently separated Army veterans was designed and selected, and a multi-method approach to survey administration was implemented to include multi-wave survey mailings, telephone followup of mail nonrespondents, and conducting a limited respondent tracing effort. Each of these survey components are described below.

---

*The views expressed in this paper are those of the author and do not necessarily reflect the official policy or position of the Department of the Army, the Army Research Institute for the Behavioral and Social Sciences, or the U.S. government.

Sample Design. An extensive sample design effort was undertaken to select a representative sample of recent Army veterans. The population of interest was defined as enlisted soldiers who separated from service between October 1981 and September 1984.* The sample was drawn using Army personnel records maintained on the FY82, FY83, and FY84 versions of the Enlisted Master File. The targeted population was further defined on the basis of their separation statuses. The survey sample was selected drawing most heavily upon young soldiers separating after completion of a successful first term of enlistment. Smaller samples were also obtained on four other groups of separatees including: first-term attritees, soldiers separating at two different points in mid-career (i.e., soldiers serving more than one term but less than ten years, and soldiers serving more than ten years but not retired), and enlisted retirees.

The distribution of sample members by separation status is presented in Table 1. As the table suggests, the group of primary interest was the sample of first-term separatees. A related priority was to be able to perform comparisons of first-term separatee and attritee responses. The samples of separated mid-careerists and retiree groups were of secondary interest in the survey effort.

Table 1

Separation Status of Army Experience Survey Sample

Separation Status	n
First-term Separatees	5,413
First-term Attritees	1,616
Soldiers with more than one term of service, but less than 10 years	601
Soldiers with 10 or more years of service, but not retired	500
Retirees	500
Unknown separation status	123
TOTAL SAMPLE	8,753

The sample design used provided a representative sample of the desired size for each of the six separation groups identified in Table 1. The individual categories of five major variables

*Excluded from the sample frame were: soldiers who separated to enter officer programs, medical retirees, soldiers who died while in service, and separating soldiers granted early release for reasons other than insufficient retainability.

were used to determine sample representativeness. These variables included race, gender, level of AFQT (Armed Forces Qualification Test), initial term of enlistment, and time elapsed since separation. Table 2 presents the 15 categories which were considered in designing the sample.

Table 2

Sampling Categories (Within Separation Groups)

Race	Gender	Initial Term of Enlistment	AFQT Level	Time Elapsed Since Separation
White	Male	2 years	Categories I and II	1 year
Black	Female	3 years	Category IIIA	2 years
			Category IIIB	
Hispanic		4 or more years	Category IV and below	3 years

A representative systematic random sample was drawn from the population frame. Then the basic sample was supplemented with additional systematic random samples to achieve the desired sample size across the sampling categories. This design resulted in a self-weighting sample.

Survey Administration. The survey administration took a multi-method approach. In general, survey procedures consisted of multiple survey mailings, respondent tracing, and telephone followup of nonrespondents. Address information was provided by the Defense Manpower Data Center (DMDC) on computer tapes. Respondent tracing procedures were employed for cases in which: (1) no address was available from DMDC, (2) survey mail was returned with no forwarding address, and (3) telephone followup was required. Each of these AES administration procedures is briefly reviewed below.

DMDC provided the complete set of Army personnel records from the Enlisted Master Files for FY82, FY83, and FY84. In addition, they performed record searches of several different military personnel files to locate address information for this population.

A sample of 8,753 veterans separating from the Army during FY82-FY84 was selected from the EMF files. DMDC was able to provide at least one address for 83% of the sample (n=7,232). In some instances, multiple addresses were provided. There were no initial working addresses for 17% of the sample (n=1,521). These cases underwent sample tracing.

A variety of sources were employed to assist in locating addresses and/or telephone numbers for sample members. Sources consulted include: NPRC/RCPAC (National Personnel Record Center/Reserve Component Personnel Administration Center), the U.S. Postal Service, the Johns Holding Company, Telematch, and Directory Assistance.

The sample was separated into three groups distinguishable by the type of tracing required: Group I had addresses provided by DMDC which were presumed valid;* Group II sample members did not have addresses provided by DMDC and required manual searches of personnel records at NPRC/RCPAC; addresses for the third group of sample members could not be located either on DMDC tapes nor at NPRC/RCPAC and were, therefore, forwarded to the Johns Holding Company (the credit bureau) for address searching. Sample members who apparently had good addresses but who never responded to the survey were forwarded to the Telephone Research Center for followup. In order to obtain telephone numbers, sample members' names and addresses were sent to Telematch and/or Directory Assistance was contacted in the city of their last known address.

Workable addresses were never located for 2% (n=166) of the AES sample. However, 14.5% (n=1,270) of the sample members were located by NPRC/RCPAC, and 5% (n=418) of the sample were located by the Johns Holding Company. Telematch confirmed addresses and provided phone numbers for 28.6% (n=2,501) of sample members. Directory Assistance was contacted when cases referred for telephone followup either had no telephone number provided by Telematch or the number provided did not reach the sample member.

Upon receipt of an address, the sequence of survey administration proceeded as follows:

- Mailing of prenotification letter;
- First-wave mailing of survey;
- Reminder/thank you postcard;
- Second-wave** mailing of survey and prenotification letter; and
- Telephone followup interviewing.

---

*Valid addresses for Group I sample members met one of the following criteria: (1) the mail was not returned by the Postal Service, or (2) mail was returned with an address correction and mail sent to the new address was not subsequently returned by the Postal Service. Mail which was not returned as undeliverable was, for the most part, considered to have reached the intended party. This assumption was tested through two mail experiments.

**In cases in which no response was received to the first survey mailing and multiple addresses were available, either from DMDC or NPRC/RCPAC, surveys were sent to all known addresses.

Survey Results. Table 3 presents the results of the survey effort by method of contact resulting in a completed survey (i.e., mail, telephone). As Table 3 indicates, a majority of completed surveys were obtained through the first mailing of the survey (62.0%). The second wave mailing elicited a fair response rate bringing in another 779 completed surveys (18.6%) of the total completes. Telephone followup provided an additional 816 completed surveys--a very successful effort particularly considering that their sample was taken from among the respondents who had already been sent four survey mailings without achieving a response.

Table 3

Sample Completion Rate by Survey Method

Mail Sample n=8,378*		
	N	% of Completes
1st Wave Mail Completes	2,601	62.0%
2nd Wave Mail Completes	779	18.6%
Telephone Followup Completes	816	19.4%
	4,196*	100.0%

*Does not include completed surveys from Telephone Only Sample, n=162.

Research Implications

The Army Experience Survey is the first known successful effort among the military services to survey veterans about their post-service experiences, and the impact of their service experience upon their civilian lives. Given schedule and budgetary constraints, the outcome of the tracing efforts and survey response rates are encouraging. Telephone followup of nonrespondents appears to be a highly effective method of increasing sample response rates and may prove to increase sample representativeness by reaching sample members less likely to complete written surveys.

USAF Spouse Survey: The Final Chapter

Major Mickey R. Dansby
Captain Karl A. Ibsen
Leadership and Management Development Center
Maxwell AFB, Al 36112-5712

The USAF Spouse Survey (AFSS) was completed by over 11,000 spouses of Air Force military and civilian personnel during the period 1982-1985. The survey was administered as a part of management consulting conducted by the Leadership and Management Development Center (LMDC). Short (1985) gives a brief history of the consulting process and describes the Organizational Assessment Package (OAP), the main consulting instrument with which the AFSS was linked. Because of its link with the OAP, the AFSS has proven to be a unique source of information on the relationship between work and family issues (Dansby, 1984; Dansby & Hightower, 1984; Flannery & Dansby, 1985). Ibsen and Austin (1983) give an overview of the purposes and history of the AFSS.

In late 1984, a decision was made to discontinue the management consulting program (by 1 Oct 86) because of manning constraints. Collection of data via the AFSS has already ceased. The purpose of the present paper is to summarize the final results of the AFSS for 1982-1985, with a breakdown of results by calendar year.

Method

The AFSS consists of 73 attitudinal and demographic items. Responses to attitudinal items range across a seven point Likert scale with a "1" indicating strong disagreement or dissatisfaction and a "7" indicating strong agreement or satisfaction. Responses to the survey were collected at numerous CONUS and overseas bases during management consulting visits solicited by commanders of Air Force organizations (usually of wing size or equivalent). All military and civilian personnel present for duty were administered the OAP in group settings. Married personnel were given the AFSS to take home to their spouses. Approximately 35% of the spouses returned the completed surveys (in sealed envelopes) to a central collection point. Subsequently, the survey responses were added to the cumulative LMDC data base of AFSS responses.

Although the total data base includes data from 30 bases (19 CONUS and 11 overseas) in nine major commands [seven CONUS (71.8%) plus USAFE (25.4%) and PACAF (2.8%)], the sample was not selected to be representative of the Air Force. Furthermore, the distribution of commands sampled varies from year to year. Therefore, the reader is advised that generalizations to the Air Force as a whole, or comparisons across year groups for trends, must be approached with caution.

Results

Table 1 provides a year-by-year and total summary for the demographic characteristics of the survey respondents. The table shows the total number of responses by year and gives demographic data as percentages of the valid responses to each item. The total number of responses may vary slightly for each item due to missing data.

Figure 1 shows the attitudinal items for the AFSS. Table 2 presents the item means, standard deviations and percent of respondents marking the item five or higher (%5+) for these items year by year and overall. For item 36 (satisfaction with the open mess) the results are given separately for spouses of officers and enlisted personnel. For items 44 and 45 (career intention), the scale has only six points; accordingly, the percentage marking four or higher is reported. Items 62 and 63 (TDY frequency and length) are primarily demographic in nature. No means or standard deviations are reported for these items (however, the table shows the %5+).

Discussion

As was mentioned previously, the AFSS was not given to a true probability sample each year, nor were the commands stratified across years. Despite these limitations, the results show a remarkable consistency from year to year. This fact, plus the broad representation of different commands and bases in the total sample, leads us to speculate that the results may be generalizable to the total population of Air Force spouses, at least insofar as broad attitudes toward the Air Force life are concerned. At the very least the data are representative of the bases at which they were collected, and the results reflect the attitudes of a significant portion of Air Force spouses.

Previous factor analyses (Dansby, 1984) resulted in the extraction of 14 factors with eigenvalues greater than one. These factors proved relatively stable over time (Dansby, 1984). We will discuss the total results of the AFSS in relation to some of the more interesting factors.

Factor 1 is a broad factor representing the spouse's identification with the Air Force. It includes items 16, 17, 19, 22, 27, 44, 45, and 71 (negative loading). Looking at the overall results for these items, we see a strong identification with the Air Force. About three quarters of the spouses are glad the member chose the Air Force as a career (item 27); a similar proportion want the military member to make the Air Force a career, if they have not done so already (item 44). Better than 60% would recommend the Air Force career (item 17) and feel involved with the Air Force lifestyle (item 16). Only about a third want the

Table 1

Demographic Summary of AFSS Data Base

Characteristic	Level	1982 n=3039	1983 n=1896	1984 n=3824	1985 n=2301	Total n=11060
Sex	Male	8.7%	11.2	9.8	9.3	9.6
	Female	91.3	88.8	90.2	90.7	90.4
Age	17-20	5.7	5.4	5.2	3.3	4.9
	21-25	20.0	19.0	25.4	18.4	21.4
	26-29	21.4	18.8	20.2	18.9	20.0
	30-34	24.0	20.4	19.8	24.4	22.0
	35-39	17.1	16.1	17.0	20.0	17.6
	40-49	9.4	15.8	10.0	12.8	11.4
	50 or more	2.4	4.5	2.4	2.2	2.7
Years Married	Less than 1	7.8	7.2	8.9	5.6	7.6
	1 but less than 4	22.7	22.3	26.4	21.6	23.7
	4 but less than 8	18.9	19.3	20.9	22.7	20.5
	8 but less than 12	20.5	15.4	15.1	16.3	16.9
	12 but less than 16	14.9	13.4	14.0	17.0	14.8
	16 but less than 20	8.5	9.6	7.8	10.0	8.7
	20 or more	6.6	12.8	6.8	6.8	7.8
Ethnic group	Indian-Alaskan	1.2	.6	1.1	1.0	1.0
	Asian-Pacific	4.1	2.1	6.8	7.6	5.4
	Black	6.3	5.5	6.6	7.9	6.6
	Hispanic	2.5	1.8	4.0	7.7	4.0
	White	83.5	88.1	79.6	73.3	80.8
	Other	2.5	1.9	1.9	2.4	2.2
Personnel Category (respondent's spouse)	Officer	17.9	23.6	23.8	24.5	23.0
	Enlisted	81.4	60.3	70.2	69.0	68.7
	Civilian	7.4	16.1	6.2	6.5	8.3
Years with the AF (respondent's spouse)	2 or less	5.9	6.7	8.4	5.3	6.8
	2-4	12.1	12.8	12.6	9.0	11.7
	4-8	17.6	18.8	23.6	18.9	20.2
	8-12	20.4	17.1	15.4	18.4	17.7
	More than 12	44.0	44.7	40.1	48.4	43.7
Time on Present Base	6 months or less	12.2	12.2	14.5	10.8	12.7
	6-18 months	28.6	27.5	32.0	30.0	29.9
	18-36 months	36.0	29.4	30.9	38.2	33.6
	More than 36 months	23.2	30.8	22.6	21.0	23.8
Housing Information	On base with member	36.7	37.8	43.5	44.4	40.9
	On another base	6.2	2.9	3.2	1.8	3.6
	Off base renting	33.9	24.7	25.9	32.9	29.4
	Off base buying	23.2	34.6	27.4	20.9	26.1
Why Living On Base?	Better schools	2.9	2.4	2.7	3.9	3.0
	Off base expensive	24.2	25.8	27.2	20.9	25.6
	Off base unavailable	2.6	1.1	1.0	1.2	1.5
Why Living Off Base?	Better schools	1.1	.8	.8	.3	.8
	No base housing	15.6	7.3	6.6	8.0	9.5
	Investment	13.1	16.0	17.3	13.5	15.2
	Not eligible/on base	6.5	11.1	8.7	7.5	8.3
	Base housing bad	12.4	13.5	13.3	16.4	13.7
	Other	12.9	13.8	10.6	13.5	12.4
Education Level	Non-high school grad	7.3	5.6	7.2	6.7	6.8
	High school grad	41.2	35.1	35.5	34.7	36.8
	2 yrs col or less	21.9	21.9	24.4	24.7	23.4
	More than 2 yrs col	14.2	17.1	16.9	17.2	16.3
	College graduate	12.2	15.6	12.7	12.5	13.0
	Masters or doctorate	3.2	4.7	3.3	4.2	3.7
Number of children at home	None	24.8	27.3	28.4	26.0	26.7
	1	24.2	25.1	26.0	24.0	24.9
	2	35.4	32.1	31.5	34.4	33.3
	3	12.5	11.3	10.9	12.0	11.7
	4-5	3.0	3.9	2.9	3.2	3.1
	6 or more	.1	.3	.3	.4	.3
Employment	Not Working					
	(don't desire)	31.6	30.9	27.7	26.2	29.0
	(not available)	19.4	23.1	19.5	24.2	21.1
	Part time work	14.0	14.8	16.8	17.3	15.8
	Active duty mil.	10.5	10.0	9.4	10.9	10.1
	Civil service	7.0	6.5	5.3	6.1	6.2
	Other	17.4	14.6	21.2	15.3	17.8
Usual Work Schedule	Not employed	50.8	54.1	47.2	50.2	50.0
	Day shift	35.3	33.7	37.4	36.1	35.9
	Other shifts	7.1	5.8	8.4	6.5	7.2
	Unstable hours/ on call	6.7	6.4	7.0	7.2	6.9
Reason for Working	Financial need	20.3	17.4	21.4	17.4	19.6
	"Extra" money	10.4	9.6	11.5	11.0	10.8
	Personal/profess. growth	16.1	17.0	17.8	18.0	17.2
	Other	2.6	2.3	2.4	3.3	2.7
Student Status	Not a student	87.0	85.0	84.9	86.3	85.8
	Full time undergrad	2.1	2.9	2.7	2.2	2.5
	Part time undergrad	6.7	7.8	9.3	7.4	7.6
	Full time grad	.5	.6	.5	.5	.5
	Part time grad	1.9	2.0	1.5	1.6	1.7
	Other	1.8	1.7	2.0	2.1	1.9
Volunteer Work	Non-volunteer	75.4	70.6	71.1	68.0	71.6
	On base vol.	12.1	11.5	12.6	16.5	13.1
	Off base vol.	7.7	12.0	10.1	8.6	9.5
	Both on and off base	4.8	5.8	6.2	6.8	5.9

BEST COPY AVAILABLE

Please indicate your agreement by choosing the phrase which best represents your attitude concerning the following statements.

1 = Strongly disagree
2 = Moderately disagree
3 = Slightly disagree
4 = Neither agree nor disagree
5 = Slightly agree
6 = Moderately agree
7 = Strongly agree

16. I feel involved with the Air Force lifestyle.

17. I would recommend an Air Force career for any young man or woman, including a son or daughter of mine.

18. My participation in base or organizational activities is essential for my spouse to achieve his/her full promotion potential in the Air Force.

19. An Air Force career has as much prestige and status as a civilian career.

20. I am interested in being informed and kept up-to-date on subjects related to the Air Force role and mission.

21. It is important for me to know about the kind of work my spouse is doing.

22. The Air Force has made considerable efforts to make service life more attractive for members and their families.

23. My spouse has to devote more time to "staying competitive" for promotion by means of service schools, college degrees, etc., than does his/her civilian counterpart.

24. My spouse has been under a lot of pressure as a result of his/her Air Force job.

25. My spouse's abilities are fully used in his/her current job.

26. My spouse has an important job.

27. I am glad my spouse chose the Air Force as a career.

28. My spouse feels positive about his/her contribution to the Air Force.

29. My spouse has to devote more time to his/her job than his/her civilian counterpart.

30. I would encourage my spouse to extend his/her military career if there were fewer moves.

31. The effect of PCS moves on family life is an important factor in my spouse's career decision.

32. Air Force leaders are sensitive to the needs of Air Force families.

533

For the various services listed below, please indicate your level of satisfaction.

1 = Extremely dissatisfied
2 = Moderately dissatisfied
3 = Slightly dissatisfied
4 = Neither satisfied nor dissatisfied
5 = Slightly satisfied
6 = Moderately satisfied
7 = Extremely satisfied

33. Base Exchange

34. Commissary

35. Military Medical Care

36. Open mess

37. Recreation center

38. Base library

39. Auto hobby shop

40. Bowling Center

41. Golf

42. Arts and Crafts

43. Child care

44. Which of the following best describes your desires for your spouse's career or employment intentions?

1. I would like my spouse to separate/terminate from the Air Force as soon as possible.
2. For the most part, I would like my spouse to not make the Air Force a career.
3. I am undecided as to my desires concerning my spouse making the Air Force a career.
4. For the most part, I would like my spouse to make the Air Force a career.
5. I would like my spouse to make the Air Force a career.
6. I would like my spouse to retire in the next 12 months.

45. Your spouse may have different career intentions than you would hope. Which of the following best describes your spouse's career or employment intentions?

1. Will separate/terminate from the Air Force as soon as possible
2. Will most likely not make the Air Force a career
3. May continue in/with the Air Force
4. Will most likely continue in/with the Air Force as a career
5. Will continue in/with the Air Force as a career
6. Planning to retire in the next 12 months

Figure 1. Attitudinal Items

BEST COPY AVAILABLE

Using the responses provided below, please indicate the extent to which you believe each of the factors listed is important in determining your spouse's career intention.

1 = Not at all
2 = To a very little extent
3 = To a little extent
4 = To a moderate extent
5 = To a fairly large extent
6 = To a great extent
7 = To a very great extent

46. Job satisfaction
47. Status and prestige
48. Rate of pay
49. Medical/dental benefits
50. Job Security
51. Retirement
52. Patriotism
53. Other

Using the same set of responses, please indicate the extent to which each of the factors listed is important in how you feel about your spouse's career intention.

54. Job satisfaction
55. Status and prestige
56. Rate of pay
57. Medical/dental benefits
58. Job security
59. Retirement
60. Patriotism
61. Other

TDY is defined as temporary military duty, and the maximum length of a TDY assignment is 179 days.

62. My spouse's job requires him/her to be TDY:

1. Less than once a year
2. Once or twice a year
3. 3 to 5 times a year
4. 6 to 9 times a year
5. 9 to 11 times a year
6. Once or twice a month
7. More than twice a month

63. How long does each TDY normally last?

1. Less than 3 days
2. More than 3 but less than 7 days
3. More than 7 but less than 14 days
4. More than 14 but less than 21 days
5. More than 21 but less than 30 days
6. More than 30 days
7. Duration varies widely

For the following items, select the most appropriate phrase and enter the corresponding number on the response sheet.

1 = Not at all
2 = To a very little extent
3 = To a little extent
4 = To a moderate extent
5 = To a fairly large extent
6 = To a great extent
7 = To a very great extent

64. To what extent does the frequency of your spouse's TDY affect your family's life?

65. To what extent do the length of your spouse's TDYs affect your family's life?

66. To what extent do you believe TDY requirements influence your spouse's career intentions?

67. To what extent do the TDY requirements of your spouse's job influence your opinion of the desirability of the Air Force lifestyle?

Below are items which relate to your spouse's job. Read each statement carefully and then decide to what extent the statement is true of your spouse's job. Indicate the extent to which the statement is true for his/her job by choosing the most appropriate phrase.

1 = Not at all
2 = To a very little extent
3 = To a little extent
4 = To a moderate extent
5 = To a fairly large extent
6 = To a great extent
7 = To a very great extent

68. To what extent do your spouse's duty hours disrupt your family life?

69. To what extent is your attitude about your spouse's job an important consideration to him/her?

70. To what extent are you proud of your spouse's job?

71. To what extent would you be happier if your spouse was doing a similar job only as a civilian?

72. To what extent would you like your spouse to change the job he/she is now doing, but remain in the Air Force?

73. To what extent do you believe that the pay and allowances earned by your spouse are in proportion to the job he/she performs?

619

Figure 1. (cont.)

Table 7

Attitudinal Summary of AESS Data Base

State-ment No.	1982 (n=1019) Mean	SD	%5+	1983 (n=1896) Mean	SD	%5+	1984 (n=1824) Mean	SD	%5+	1985 (n=2060) Mean	SD	%5+	Total (n=10819) Mean	SD	%5+
16	4.65	1.91	619	4.71	1.90	62	4.81	1.87	66	4.88	1.89	65	4.76	1.88	63
17	4.54	1.96	56	4.86	1.88	61	4.92	1.87	65	4.77	1.93	61	4.77	1.91	61
18	2.97	1.99	24	3.20	2.03	29	3.51	2.07	34	3.44	2.09	32	3.29	2.05	30
19	4.33	2.14	52	4.63	2.11	56	4.73	2.09	60	4.47	2.15	54	4.55	2.12	56
20	5.56	1.58	78	5.70	1.44	81	5.66	1.53	80	5.58	1.59	78	5.62	1.54	79
21	6.16	1.31	89	6.25	1.20	92	6.25	1.22	91	6.10	1.34	89	6.18	1.27	90
22	4.25	1.94	55	4.37	1.92	58	4.44	1.91	59	4.41	1.95	58	4.36	1.92	58
23	4.79	1.79	56	4.81	1.79	56	5.03	1.76	61	5.05	1.79	61	4.92	1.78	59
24	5.28	1.78	73	5.21	1.78	73	5.35	1.74	75	5.42	1.73	75	5.32	1.75	74
25	4.25	2.28	52	4.32	2.27	53	4.40	2.24	55	4.38	2.26	54	4.33	2.25	53
26	6.00	1.44	86	6.04	1.41	87	6.09	1.40	88	6.14	1.40	88	6.06	1.41	87
27	5.27	1.79	68	5.52	1.67	74	5.58	1.67	76	5.46	1.75	73	5.45	1.72	73
28	5.49	1.73	76	5.76	1.58	82	5.74	1.60	82	5.69	1.66	81	5.66	1.65	80
29	5.49	1.74	72	5.35	1.77	67	5.53	1.69	73	5.66	1.67	73	5.51	1.71	71
30	4.65	1.97	50	4.64	1.93	48	4.73	1.96	51	4.61	2.02	47	4.66	1.96	50
31	5.05	1.89	62	5.00	1.89	61	5.07	1.86	62	5.06	1.95	63	5.04	1.88	62
32	3.39	1.90	35	3.61	1.92	38	3.62	1.90	39	3.54	2.00	37	3.52	1.91	37
33	3.83	1.90	45	4.11	1.85	51	4.24	1.84	54	4.01	1.92	48	4.05	1.87	50
34	4.51	1.87	60	5.05	1.70	70	4.77	1.86	66	4.62	1.86	61	4.70	1.84	64
35	4.26	2.03	53	4.33	2.01	54	4.11	2.06	50	4.25	2.09	52	4.20	2.04	52
36 (Off)	4.05	1.66	42	4.19	1.71	42	3.64	1.73	31	3.87	1.73	37	3.89	1.72	37
36 (Enl)	3.99	1.52	29	3.98	1.48	28	4.21	1.50	34	4.00	1.58	32	4.07	1.52	32
37	4.39	1.30	38	4.25	1.20	29	4.52	1.26	39	4.45	1.38	39	4.42	1.28	37
38	5.06	1.33	60	5.12	1.34	62	5.10	1.32	61	5.18	1.38	65	5.10	1.33	62
39	4.39	1.14	29	4.40	1.07	28	4.54	1.17	33	4.36	1.22	27	4.43	1.15	30
40	4.90	1.40	53	4.76	1.30	49	4.75	1.33	48	4.85	1.39	51	4.81	1.35	50
41	4.25	1.10	19	4.31	1.14	21	4.49	1.14	27	4.32	1.22	22	4.35	1.14	21
42	4.45	1.20	34	4.52	1.18	34	4.43	1.24	32	4.40	1.30	32	4.43	1.22	33
43	3.88	1.59	26	3.94	1.60	24	4.12	1.53	30	3.93	1.63	24	3.97	1.57	26
44	3.97	1.26	70 (4+)	4.08	1.20	71 (4+)	4.08	1.21	73 (4+)	4.13	1.37	71 (4+)	4.04	1.23	72 (4+)
45	4.08	1.23	72 (4+)	4.13	1.17	75 (4+)	4.12	1.21	74 (4+)	4.21	1.30	76 (4+)	4.12	1.21	74 (4+)
46	5.75	1.49	81	5.85	1.37	84	5.78	1.41	82	5.76	1.52	81	5.77	1.44	82
47	4.42	1.69	47	4.46	1.64	48	4.54	1.65	50	4.52	1.72	50	4.48	1.66	49
48	5.26	1.62	69	5.33	1.51	72	5.24	1.59	70	5.20	1.67	67	5.24	1.60	70
49	5.01	1.64	64	4.91	1.69	62	5.12	1.62	66	5.10	1.71	65	5.04	1.65	65
50	5.58	1.45	79	5.72	1.40	81	5.66	1.41	81	5.61	1.51	78	5.63	1.44	80
51	5.66	1.49	80	5.69	1.47	80	5.71	1.48	80	5.76	1.50	80	5.70	1.48	80
52	4.97	1.68	61	5.08	1.62	63	5.29	1.60	70	5.30	1.63	68	5.16	1.63	66
53	4.22	1.96	42	4.29	1.91	43	4.34	1.97	46	4.60	2.09	46	4.31	1.95	44
54	5.81	1.52	82	5.93	1.42	84	5.87	1.46	83	5.85	1.53	82	5.85	1.48	83
55	4.47	1.78	50	4.55	1.76	51	4.62	1.72	53	4.60	1.78	52	4.55	1.74	52
56	5.39	1.64	73	5.47	1.53	75	5.44	1.55	74	5.43	1.64	73	5.42	1.58	74
57	5.25	1.71	70	5.20	1.74	69	5.39	1.64	73	5.39	1.71	73	5.31	1.69	71
58	5.23	1.44	82	5.88	1.35	84	5.81	1.41	83	5.75	1.49	81	5.78	1.42	82
59	5.66	1.52	79	5.68	1.50	80	5.74	1.49	81	5.79	1.53	80	5.71	1.50	80
60	4.74	1.80	55	4.84	1.75	57	5.05	1.73	63	5.08	1.78	62	4.93	1.76	60
61	4.10	2.02	40	4.19	1.97	41	4.25	2.03	44	4.50	2.21	43	4.20	2.01	42
62	-	-	6	-	-	12	-	-	7	-	-	11	-	-	8
63	-	-	40	-	-	24	-	-	34	-	-	36	-	-	34
64	3.27	2.05	27	3.15	1.95	23	3.27	1.97	44	3.49	2.08	27	3.28	1.99	26
65	3.64	2.13	34	3.49	2.04	31	3.64	2.06	33	3.86	2.11	36	3.65	2.08	34
66	3.30	2.06	30	3.21	1.98	27	3.41	2.04	31	3.45	2.09	29	3.34	2.03	29
67	3.51	2.18	32	3.41	2.08	30	3.53	2.09	32	3.72	2.16	34	3.53	2.11	32
68	3.33	1.97	27	3.05	1.93	22	3.21	1.93	25	3.47	2.12	29	3.26	1.97	26
69	5.13	1.58	66	5.15	1.59	68	5.24	1.57	71	5.26	1.61	69	5.19	1.58	69
70	5.60	1.49	78	5.76	1.41	81	5.82	1.40	82	5.81	1.47	82	5.74	1.44	81
71	3.85	2.12	37	3.56	2.12	32	3.74	2.06	34	4.06	2.18	38	3.79	2.09	36
72	3.17	2.22	28	3.09	2.18	27	3.13	2.18	27	3.37	2.31	29	3.16	2.20	28
73	3.19	1.60	18	3.21	1.59	20	3.26	1.60	19	3.31	1.87	18	3.21	1.60	19

BEST COPY AVAILABLE

military member to leave the Air Force and pursue a similar career as a civilian (item 71). Almost 60% perceive the Air Force as going to considerable effort to make service life more attractive (item 22). These results show fairly strong spousal support for the Air Force.

The second factor represents job benefits as a retention influence. Items 49, 50, 51, 57, 58, and 59 load with this factor. Since items 49, 50, and 51 are highly correlated with items 57, 58, and 59 (respectively), we will discuss only the first set of items. Almost two thirds of the spouses see medical benefits as a determinant of career intention to at least a "fairly large extent." About 80% of the spouses indicate job security and the retirement program are strong determinants of career intention.

TDY attitudes (items 64-67; Factor 3) show a relatively small impact on family life and career intentions. Less than a third of the spouses say TDY length or frequency influences their family life (items 64 and 65), or the member's career intentions (item 66) and opinion of the desirability of Air Force life (item 67), to a "fairly large extent" or more. On the other hand, results for Factor 4 (satisfaction/prestige as a career influence) show the spouses believe job satisfaction (item 46), status and prestige (item 47), and pay (item 48) are important career determinants. Over 80% rate satisfaction (item 46) as important to a "fairly large extent" or more; the corresponding proportions for status (item 47) and pay (item 48) are 50% and 70%.

Items (37-42), loading on the fifth factor, measure satisfaction with recreation facilities. This factor will not be discussed, since satisfaction varies from base to base and depending on frequency of use. Factor 6 (identification with job) is interesting. Most of the spouses (over 80%) are proud of the member's job (item 70) to at least a "fairly large extent"; few (less than 30%) want the member to change jobs within the Air Force (item 72). A resounding majority (nearly 90%) agree that the member's job is important (item 26); however, only a little better than half agree that the member's abilities are being fully used on the job (item 25). Most spouses (about 80%) agree that the member feels positive about his or her contribution to the Air Force (item 28).

Factor 7 reflects attitudes on basic services (items 33-36 and 43). For this factor, we note that the commissary (item 34) receives the highest ratings of satisfaction, followed by the base exchange and medical care, and finally the open mess and child care services. Pressure from the job (Factor 8) appears to be a significant concern for the spouses. About three fourths agree that the member has been under a lot of pressure as a result of his or her Air Force job (item 24). While only a fourth of the spouses respond that the member's job disrupts family life to at least a "fairly large extent" (item 68), over 70% agree that the member devotes more time to the job than his or her civilian counterpart (item 28). About 60% agree that the member must devote more time to "staying competitive" than a civilian counterpart (item 23).

Factor 9 (other influences on career decisions) yields little meaningful information. Factor 10 (patriotism as a career influence; items 52 and 60) is rated high by the spouses. About two thirds believe patriotism influences the member's career intention to a "fairly large extent" or more. The eleventh factor (desire for information about the member's job; items 20 and 21) has importance for Air Force commanders and supervisors. Over 80% of the spouses agree that they are interested in being kept up to date on the Air Force role and mission (item 20), and over 90% want to know about the work the member is doing (item 21).

Factors 12-14 will not be discussed (they are addressed in Dansby, 1984). Two items that have considerable importance for Air Force leaders should be noted, however. Over 60% agree that PCS moves have an important effect on family life and retention (item 31). Finally, less than 40% of the spouses agree that Air Force leaders are sensitive to family needs (item 32).

The results of the AFSS present some challenges for Air Force leaders. Clearly, the spouses like and support the Air Force; but just as clearly, they expect the Air Force to be responsive to the needs of the family. Open communication between families and Air Force officials needs to be maintained so leaders can be responsive. Further research with the AFSS, or a similar instrument, could serve to keep Air Force leaders in touch with the "pulse" of Air Force family life. Response to family needs may be critical in retaining a dedicated, experienced, quality force.

References

Dansby, M. R. (1984). A proposal for the revision of the U.S. Air Force Spouse Survey (Report No. 84-0645). Maxwell AFB, Al: Air Command and Staff College.

Dansby, M. R. & Hightower, J. M. (1984). Family and work in the Air Force. Proceedings, Psychology in the Department of Defense, Ninth Symposium, 455-459.

Flannery, P. A. & Dansby, M. R. (1985). USAF Family Survey: A revision of the USAF Spouse Survey (Report No. LMDC-TR-85-3). Maxwell AFB, Al: Leadership and Management Development Center.

Ibsen, K. A. & Austin, J. S. (1983). Initial development of the USAF Spouse Survey. Proceedings of the 25th Annual Conference of the Military Testing Association, 633-638.

Short, L. O. (1985). The United States Air Force Organizational Assessment Package (Report No. LMDC-TR-85-2). Maxwell AFB, Al: Leadership and Management Development Center.

BEST COPY AVAILABLE

PREDICTORS OF PROPENSITY FOR CONTINUING EDUCATION AMONG ARMY CHAPLAINS

J. ERIC PIERCE, Ph.D., Chapel of the Four Chaplains
LAWRENCE A. GOLDMAN, Ph.D., USA Soldier Support Center-NCR

Background. Army chaplains, upon induction, have taken a minimum of 90 semester hours of academic work beyond the college level. As it requires roughly three years of full-time post graduate study to accumulate these 90 semester hours, one might expect chaplains to be individuals who are positively disposed toward continuing education. Chaplains, however, display varying amounts of positive attitudes toward continuing education. Personal experience, as well as responses to the instrument used in the present research, indicate that some chaplains are far more likely than others to seek or accept opportunities for continuing professional education. While there may be many contributing factors which would help to explain such a variety in attitudes, it would be most beneficial for the Chaplain Branch of the U.S. Army to be able to isolate some of the most sensitive predictors of propensity to pursue continuing education. For the purpose of this study, "continuing education" was defined in an Army-wide survey of chaplains conducted in 1984 "as a carefully guided reading program, a directed course of study at home or away from home, formal classwork, or seminary workshops of two or more weeks duration".

The Chaplain Branch administers a professional educational system of its own, as do other branches of the Army. This system includes the offering of courses at the U.S. Army Chaplain Center and School at Fort Monmouth, NJ, most of which are mandatory, as well as selection of certain officers to participate in The Army Educational Review Board program. Under this program, and because of special professional needs of the Branch and the services it is called upon to offer the Army community, the Army selects chaplains for special training at civilian institutions. A reliable set of predictors of propensity among chaplains to seek continuing education would facilitate this process, especially if future studies were to establish a positive correlation between prediction scores and successful academic performance. For the present study, we focused upon isolating the most reliable predictors from which such scores could be derived.

Major Variables and Models. The methodology used for this study consisted of deriving concepts from extant literature about continuing education among professionals and from pilot interviews with chaplains in the field and at the U.S. Army Chaplain Center and School. From these concepts, measurable variables were derived, and a questionnaire was constructed. The original study attempted to explore two research

questions: "What tasks are characteristic of the various duty positions of chaplains?", and "What is the nature of the relationship between the chaplaincy and the continuing education experience?". It was from exploration of the latter question that we discovered certain predictors of propensity toward involvement in continuing education among chaplains by isolating common factors among respondents who expressed a high degree of likelihood that they would seek further education.

Chaplains are clergy persons. The exercise of their profession involves a particular set of skills that can be taught and practiced under supervision. It entails an institutional setting through which service is rendered, and a formalized ethic of service. A minimum level of training for admission to the profession (seminary) is followed for many in the profession by continuing education and training to hone and preserve old skills, and to acquire new ones. The beliefs, attitudes and intentions of chaplains and other clergy are obviously among the behavioral determinants of seeking and participating in continuing education. Fishbein and Ajzen (1975) presented a theoretical framework for examining the relationships among these variables. They demonstrated that belief about an object influenced attitude and that, in turn, influenced behavior. Groteleuschen and Caulley (1977) applied the framework to continuing education as the type of behavior influenced. Consequently, they evolved a predictor model which, after some revision of its statistical formula, was useful in the present study. Bonn (1974) had devised previously a specific predictor model for participation by clergy in continuing education.

While income and previous educational experience are generally accepted as powerful contributing factors toward participation in continuing education, they were not expected to prove strong variables among chaplains for the following reasons:

1. Uniform minimum education requirement of 90 semester hours of graduate study for admission to the Army Chaplain Branch.

2. Competitive screening by most large denominations of chaplain candidates tending to promote admission of highly qualified persons.

3. Uniform pay for chaplains, by rank.

4. Government financing and mandating of some continuing education.

5. A professional value system favorably disposed toward continuing education.

Because, then, of a given degree of uniformity in so many areas, it was decided to concentrate upon beliefs, attitudes and intentions of chaplains regarding continuing education. To this end, the Groteleuschen and Caulley model was useful. The model is stated mathematically as:

$B \sim I = w_1 (A) + w_2 (SSN) + w_3 (SPN)$, where

B = The behavior in question,
I = Intention to perform behavior B,
A = Attitude toward performing behavior B,
SSN = Subjective Social Norm, and
SPN = Subjective Personal Norm, with w_1, w_2, and w_3 being empirically derived weights. Questionnaire items were provided in the model from which scores could be derived for the variables.

The Subjective Social Norm and Subjective Personal Norm were derived in the model from scores on pairs of questionnaire item indicators relating to belief and motivation. The impact of social interaction can be seen by Groteleuschen and Caulley's use of these two concepts as factors in the prediction of intention to participate in continuing education. The Subjective Social Norm relates to the actor's perception of the reaction of significant others to his participation. The Subjective Personal Norm is concerned with the actor's own beliefs about such participation. The Symbolic Interactionists would insist upon combining these two "norms" in any consideration of predictors of intention, and Groteleuschen and Caulley have done so by their formula.

Groteleuschen and Caulley had suggested that the weights in their formula be determined by running a multiple regression analysis for an entire group of respondents, using I as the criterion and A, SSN, and SPN as the three predictors. The standardized regression coefficient of each component would serve as its weight in utilizing the formula for determining I. To eliminate the problem of using I before its value is computed, we used partial correlation coefficients of each of the three components. The obtained coefficients were then used as the weights for the components.

Aside from the Bonn Predictor model and the Groteleuschen and Caulley model, we utilized scores from the following (each representing a composite of several indicators):

1) Stress and General Dissatisfaction Test
2) Perceived Barriers to Participation in Continuing Education
3) Felt Need for Continuing Education
4) Planned Participation in Continuing Education
5) Perceived Adequacy of the Respondents' Training

6) Willingness to Participate in Continuing Education
7) Status Inconsistency Test

Findings. There was a very strong relationship between present or recent participation and future participation in a deliberately planned program of continuing education experience. Ninety-five percent of chaplain respondents who were currently or recently (within 12 months) involved in a continuing educational experience expressed that they were "somewhat likely" or "very likely" to seek further education. Of those chaplains not currently or recently involved in such an experience, only 73 percent were either "somewhat likely" or "very likely" that they would be so involved in the future. Thus, past or present participation appeared to be an indicator of future participation.

A stepwise multiple regression analysis was conducted to isolate the best predictors of participation in continuing education. The seven attitudinal variables and the two predictor models comprised the independent measures. Responses ("Yes" or "No") to the question, "Are you presently participating or have you in the past 12 months participated in a deliberately planned program of continuing education?", constituted one dependent measure. Responses on a 5 point scale, ranging from "Not likely at all" to "Very likely" for the question, "In the next 5 years, how likely is it that you will participate in a deliberately planned program of continuing education?", comprised the other dependent measure. As shown in Table 1, the significant indicators of participation, in order of significance, were: The Planned Participation Test, the Groteleuschen and Caulley model, the Status Inconsistency Test, and the Perceived Barriers Test.

As shown in Table 2, the Planned Participation Test, the best independent measure of participation in current or recent continuing education, was also identified as the best predictor of perceived future participation in such a program. The simple r of .80 between the Planned Participation Test scores and perceived future participation was highly significant. Felt Need for Continuing Education was also a statistically significant predictor of future educational propensity, but to a lower degree than the Planned Participation Test.

Table 3 shows the inter-correlations of the five independent measures identified as significant predictors of the two dependent measures.

TABLE 1 - SIGNIFICANT MEASURES OF CURRENT (OR RECENT) PARTICIPATION IN A PROGRAM OF CONTINUING EDUCATION

Independent Measure	Multiple R	R^2	R^2 Change	Simple r*
Planned Participation Test	.318	.101	.101	-.32
Groteleuschen and Caulley Predictor Model	.342	.118	.017	-.22
Status Inconsistency Test	.354	.125	.007	-.05
Barriers to Participation in Continuing Education	.363	.132	.007	-.21

*Based on a dichotomous scale where "1" = yes and "2" = no

* *

TABLE 2 - SIGNIFICANT MEASURES OF FUTURE PARTICIPATION IN A PROGRAM OF CONTINUING EDUCATION

Independent Measure	Multiple R	R^2	R^2 Change	Simple r
Planned Participation Test	.801	.641	.641	.80
Felt Need for Continuing Education	.817	.668	.027	.13

* *

TABLE 3 - INTER-CORRELATION MATRIX OF SIGNIFICANT INDEPENDENT MEASURES OF CURRENT (OR RECENT) AND FUTURE PARTICIPATION IN A PROGRAM OF CONTINUING EDUCATION

	PLANPART[1]	GROTCAUL[2]	STINCTEST[3]	BARRPART[4]	FELTNEED[5]
PLANPART[1]	1.00	.35	-.07	-.32	.06
GROTCAUL[2]		1.00	-.10	-.38	.13
STINCTEST[3]			1.00	.08	.00*
BARRPART[4]				1.00	-.03*
FELTNEED[5]					1.00

1 Planned Participation Test

2 Groteleuschen & Caulley Predictor Model

3 Status Inconsistency Test

4 Barriers to Participation in Continuing Education

5 Felt Need for Continuing Education

* Not statistically significant from zero.

Conclusions. Those variables which proved to be most sensitive and useful as predictors were the five identified previously. The Groteleuschen and Caulley model, although living up to our expectations, was not quite as sensitive as the Planned Participation Test. Still, it should be retained for use in future studies. The Bonn Predictor model and three of the composite variables used in the present study (Willingness to Participate, the Stress and Dissatisfaction Test, and Perceived Adequacy of Training) proved to be less sensitve and therefore were not retained.

Useful future studies to isolate reliable predictors would include a longitudinal study to discover whether or not respondents who perceive a strong likelihood of participation in a continuing education program actually do participate within five years. Another useful study would be to track the academic performance of respondents in continuing education programs subsequent to administering the predictor tests. A high positive correlation between test scores and grades would reinforce the predictive value of the tests.

REFERENCES

Bonn, Robert L. Continuing Education Participants: Who, How Many, Types of Programs. Richmond: SACEM, 1974

Fishbein, M. and Ajzen. Beliefs, Attitudes, Intention and Behavior. Reading, Mass: Addison - Wesley, 1975.

Groteleuschen, A. and Caulley, D. "A Model for Studying Determinants of Intention to Participate in Continuing Professional Education". Adult Education, 28 (1977), pp. 22-37.

Composite Ratings as a Performance Criterion

Jonathan Warren
Consultant

Peter Newton
National Security Agency

John Bondaruk
National Security Agency

A continuing issue in studies of job performance is establishing acceptable, useful criteria. Whether the purpose of a study is to evaluate selection or promotion procedures, or training effectiveness, or to compare different classes of employees, the usefulness of the results depends heavily on the accuracy of the criteria. In a long history of attempts to improve performance criteria, behaviorally anchored rating scales and behavioral observation scales are probably the most popular current choices. Their use is accompanied by a concern for specifying and rating separately the various components of performance important in any given job. The separate ratings may or may not then be integrated into a composite global rating. The gains expected from being more precise in what we attempt to assess have been disappointing. We may do better by being more modest in our attempts to pin down a criterion with great precision while paying closer attention to some of the mundane aspects of human judgment. The rest of this paper describes such an effort.

In a study of selection procedures in a large government organization, a composite rating procedure for assessing job performance was devised with attention to the practicalities of recording accurate, thoughtful human judgments. First, since large numbers of ratings were required, the demands on the raters in time and effort should be reasonable. The process should be quick and easy to avoid rater fatigue and loss of interest. Second, the judgments asked for should be realistic, understandable (within the frame of reference of the employees rather than psychologists), clearly related to the job being assessed, and should refer to performance normally observable by the raters. Third, the ratings should permit the application of multiple perspectives. The same qualities are assessed, perhaps using the same form, but the judgments are made from different assumptions, or for different stated purposes.

The procedure used called for the raters to make three judgments about each of from six to twenty employees whose work they knew. All three judgments referred to general or global performance in a single job for which a detailed job description had recently been established and which was well known by the raters. An initial list was given each rater of up to 20

employees he or she had recently supervised. Before starting the rating process, each rater deleted the names of persons he or she felt unable to rate and added others to a maximum of 20. The average number of persons rated was about eight.

The instructions asked the raters to assume they were in a newly formed organization in charge of hiring a number of new employees for the job being rated. The persons hired were to be productive immediately without further training. The raters were instructed to identify the three persons on their list they would hire first. Then they indicated the best person among those three, and then the three next best from among the persons not yet selected.

As long as at least six names were on the list, this process selected the single most effective employee, two who were next best, and three who ranked just below them. The remaining names constituted a fourth group who trailed the others. Thus every person on the list was ranked in terms of overall productivity in the job, but without fine distinctions except at the top of the order. Avoiding those fine distinctions, which are often difficult to make and rarely informative, is one way the process was simplified and shortened without sacrificing good information.

For the second task, the raters were informed that they were to select ten persons for their new organization from a pool of about 100 applicants that included the persons on their list. They were to identify any of the listed employees the raters thought were so competent that they would hire them immediately without bothering to check the qualifications of the other applicants. This judgment, which is made fairly quickly, indicates whether any or all of the persons rated are unusually competent.

Finally, for their third task, the raters indicated on a five-point scale their absolute judgment of the overall capability of each listed employee. The five scale points were labeled from "Barely proficient" to "Among the best I've seen," with similar descriptions for the three intermediate points.

The use of three kinds of ratings kept some variability in them, balancing some of the advantages and disadvantages of relative and absolute judgments. The first step forced the raters to discriminate among the employees at four levels of relative capability. The instructions for the absolute rating scale informed the raters that they were not required to use the full range of the scale, that they could rate everyone at the top, the bottom, or anywhere else. Yet the discriminations they had just been forced to make in ranking the employees no doubt imposed some tendency to use more than a limited segment of the scale.

The variability of the absolute ratings was quite good, with

acceptable proportions of ratings in the lowest levels. On the five-point absolute scale, the percentages of ratings from lowest to highest were 3, 10, 28, 37, and 22. From lowest to highest among the four levels of relative ranks, the percentages were 54, 22, 16, and 8. The ranks thus discriminated best at the upper end of the distribution of performance, while the absolute ratings discriminated well throughout the range but best at the lower end. Slightly more than a third of the employees (37 percent) were judged capable enough to be hired immediately, providing a coarse but quick confirmatory indicator of the absolute level of performance.

About 700 supervisors judged a total of more than 5,000 employees who represented nine different jobs in five job families, with at least 240 persons in each job. Almost 2,000 employees were judged by at least two supervisors; almost 400 had been judged by four.

To estimate the reliability of the judgments, the employees were grouped according to whether they had been rated by two, three, or four supervisors. In a table with employees as rows and ratings as columns, the placement of ratings in the rows was arbitrary. Thus one of the ratings of someone who had been rated twice could appear in either the first or second column. A rating for someone rated three times could appear in any of the first three columns. No one was rated more than once by the same rater. The four columns therefore represented four independent sets of ratings.

The level of agreement between all six possible pairs of columns was calculated using gamma, a measure of agreement between two sets of ranks. Median gammas for the absolute ratings and the relative ranks were respectively .47 and .37. This procedure includes error attributable to the different judges rating each employee, which reduces the levels of interjudge agreement from those that would be reached if the same supervisors had judged all the employees, a condition not feasible with a large number of employees. For six different job families, the median coefficient between absolute ratings and relative ranks given by the same rater was .60.

With large numbers of employees judged by several raters and each rater providing three separate judgments, a number of composite ratings can be formed. A composite scale formed by combining the three ratings by a single supervisor showed an intraclass correlation coefficient of .75. The scale had a range of 2 to 12, a mean of 8.1, and a standard deviation of 1.8. Combining all three judgments by two sets of supervisors produced a scale with a greater range and about the same reliability, .77. Other composites produced reliabilities of .70 to .75, but all required judgments by more than one supervisor. The three different judgments by the same supervisor matched the reliability of the best scale formed by the judgments of more than one rater.

The criteria were used in a validity check of a battery of selection tests for employment. The predictors were tests of cognitive skills that formed five factors--verbal ability, reasoning, clerical ability, spatial perception, and knowledge of science. For many of the employees, the tests had been taken a number of years before the judgments of job performance were made. Within the different job categories, multiple correlation coefficients between optimal combinations of predictors and multiple criteria ranged from about .20 to .45 with a median of .28.

The validity coefficients for composite criteria formed from the judgments of two, three, and four supervisors were .24, .26, and .34 respectively. The numbers of cases in the groups of two, three, and four supervisors were respectively about 2,000, 800, and 400. As expected, combining the judgments of more than one supervisor apparently increases the validity of the criterion, although the gain may not be appreciable.

The best composite scale in terms of reliability and validity was the six-element composite formed from three judgments by each of two judges. Virtually on its heels, though, with respect to both reliability and validity, was the single-judge scale formed from the three ratings. The probable tendency of judges to be moe careful in their absolute ratings after having been forced to rank the persons being rated, the greater sensitivity of the ranks at the upper end of the scale and of the ratings at the lower end of the scale, and the probable corrective influence of the binary decision to hire a person immediately all may contribute to the usefulness of the single-rater composite scale. Its ease of administration is another important element. All three ratings of as many as 20 persons could be made in about 15 minutes. The different perspectives the three judgments require may provide a useful, global criterion of job performance.

MAXIMIZING CRITERION VARIANCE IN VALIDATION RESEARCH: IS THIS ALWAYS BEST?

Captain Robert J. Angus
and
Major Reginald T. Ellis

Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario, Canada

The Canadian Forces Personnel Applied Research Unit (CFPARU) is responsible for conducting a continuing research program to maintain a reliable and valid set of selection tests for use in the Canadian Forces (CF). The selection test battery currently in use is the Canadian Forces Classification Battery (CFCB), which was implemented operationally in 1981. The technical training courses, for which this battery is used to select students, are continually changing; consequently the CFCB must be closely monitored to ensure that the best students continue to be selected to fill recruiting vacancies.

Criterion Data Collection System

Initial CFCB validation studies were conducted using training performance criterion data collected specifically for that project. Recognized limitations in the quality, consistency and comprehensiveness of these one-shot data led to a decision to establish an on-going data collection system, which would start to generate data at the same time as the operational implementation of the new battery. This Criterion Data (CD) collection system was set up to support periodic studies to revalidate CFCB standards against basic trades training performance for over 60 entry level occupations.

The CD collection system provided for the acquisition of standard format training performance measures on every graduate from CF basic trades training courses (Ellis & Saudino, 1980; McMenemy, Amyot, & Enkurs, 1984). Implicit in the design of this system were two assumptions. The first was that maximizing variance in the criterion measures would yield better estimates of validity. The second was that the capacity to identify, and exclude from the analysis, those trainees who were unsuccessful for reasons unrelated to ability (i.e., motivational failures) would similarly yield better validity estimates. Thus, steps were taken to supplement readily available student computer records consisting of pass/fail information, with more precise performance

---

The views and opinions expressed in this paper are those of the authors and not necessarily those of the Department of National Defence.

measures for both the passing group (which almost invariably represents more than 90% of the sample) and for the non-successful group. Supplementary information for the passing group consisted of performance measures such as class standing or course grade. For the non-successful group, supplementary information was aimed at providing a means of discriminating between failures related to lack of ability and those related to lack of motivation.

The latter requirement was addressed by devising a system of disposition and reason-for-disposition codes for use in conjunction with automated individual training records. This required training authorities to denote, within the automated training record, the disposition action ("release" from the CF, "reassignment" to a different trade, or "recourse" to a later course serial), and the reason for non-success (lack of ability, lack of motivation, medical problems, etc.) for each non-passing trainee.

Validation Studies

Based on this data collection system, a standardized approach to constructing data sets and conducting validity analyses was established to cycle through revalidation studies on various trades as the need to do so was identified and/or as sufficient criterion data accumulated (Miller & Angus, 1985). The validation model involves the identification of the best CFCB single or composite predictor on the basis of a comparative correlational analysis. Included in the model is a procedure for selecting an appropriate cut-off score on the basis of a standard contingency table reflecting success rates associated with an array of selection ratios.

During the first of a series of revalidation studies based on the new model (Miller & Angus, 1984), it was determined that the best CFCB predictor for the trade under study yielded a validity coefficient of .31 (corrected for restriction of range). In constructing a selection ratio versus pass-rate contingency table, it was expected that as the selection cut-off score increased, the "efficiency" of the selection process would also increase (i.e., the percentage of students above the cut-off who passed the course would rise in a similar fashion). However, this pattern was not observed in the contingency table for this predictive validity study. In fact, it was observed that the success rate (selection efficiency) remained constant irrespective of the selection ratio used. Further investigation of data from the 18 students who were unsuccessful during this training revealed that nine of these trainees had achieved predictor scores above the 40th percentile. Having already identified the most valid predictor of training performance for this particular trade, it was inconsistent that students with predictor test scores in the upper part of the distribution should perform so poorly during training, if their training problems were truly ability-related. In spite of the fact that those nine students had much higher selection test scores than was necessary to meet the minimum selection standard for this course (a percentile score of 10), they failed to successfully complete the training and were eventually classified as ability-related failures.

Student "Reason for Failure" Misclassification

It is apparent that some or all of these nine students may not have failed for ability-related reasons. Rather, it seems more likely that they were incorrectly classified as ability-related failures, and should have been assigned to a motivation-related failure category. If they had been classified as motivation-related failures, their scores would not have been included in these analyses since only those students who are successful, or are unsuccessful for reasons attributable to ability, are to be included in the analyses. To examine this possible misclassification issue, the total sample of trainees was rank-ordered by predictor test scores, and those unsuccessful trainees falling above the median for this whole group were identified as "potentially misclassified". Further analyses were then conducted to determine the validity coefficients for the following sub-groups of students:

a. all those who had passed the course;

b. all those who had passed the course less the nine potentially misclassified students; and,

c. the group of 18 students who failed the course.

It was hypothesized that if misclassification had occurred, then the following results could be expected from these analyses:

a. the validity coefficient for the group who passed the course should be greater than the validity coefficient for the full group of trainees;

b. the validity coefficient for the full group less the potentially misclassified failures should be even greater than that of the passing group alone;

c. the validity coefficient for the group of unsuccessful trainees should be very low since half of the group are students that are potentially misclassified.

The results of these analyses are presented in Table 1. Examination of this table shows that the validity coefficients are completely consistent with the hypothesized relationships. Inclusion of data from students who were unsuccessful on the course for ostensibly ability-related reasons, instead of contributing positively to the precision of the validity analysis, was detrimental to its overall accuracy. This fact is demonstrated by an increase in the validity coefficient from .31 to .37 as a direct result of the exclusion of records pertaining to failed trainees, and an additional increase to .39 when only the potentially misclassified (failed) trainees are excluded. Further, the validity coefficient for the unsuccessful group of students shows that not only was the correlation between their scores and performance on basic trades training not significiantly different from zero, but the correlation was negative, or in the OPPOSITE direction to what would be expected. In other words, the inclusion of scores for misclassified failed students erroneously and significantly decreasing the validity coefficient.

Table 1

An Examination of the Change in
Validity Coefficients as a Result of Separating
Successful and Unsuccessful Trainees

Group	n	Corrected Validity Coefficient
All trainees	195	r = .31
Successful trainees only	177	r = .37
All trainees less the misclassified group	186	r = .39
Unsuccessful trainees only	18	r = -.14*

Note: All correlations are significant (at the .01 level) except the one designated by an asterisk.

Reasons for Misclassification

Misclassification in this situation is understandable if one considers some of the circumstances specific to the CF trades training system. The CF operates on a one person/one job basis, with very strict and finite limits imposed on the numbers of personnel permitted within each rank level in each trade. A recruit is therefore assigned to a single, specific trade at the point of enrolment, and only into a trade for which a specified training vacancy is open at that time. Recruits often don't get their "first choice" of trade, but may agree to enrolment in a less preferred trade in order to gain initial entry to the CF.

However, the opportunities for post-enrolment reassignment to another trade are relatively limited. If a trainee discovers, once trades training has commenced, that the originally assigned trade is not a vocation in which s/he can be happy or satisfied, there are few options open in terms of changing trades. An outright request for reassignment to another trade is in most cases not actioned, and until recently often resulted in the trainee being released if s/he was adamant about leaving the originally assigned trade. This was intended to prevent students from requesting reassignment on a "whim" if the training in their present trade was at a particularly demanding or stressful point. Early release from the CF has not, in recent times, been an attractive option with young trainees, who face the subsequent prospect of high civilian unemployment rates.

The CF does try to accommodate those trainees who do their best and work hard during trades training, but are unable (perhaps through a lack of aptitude) to successfully complete the course. These trainees are quite often recommended, by their course instructors, for reassignment to another CF trade where they might have a better chance of success. Consequently, it can be seen that trainees who appear to their instructors to have the ability to succeed, but are simply not positively motivated to

do so, are generally not considered good risks for retention in the CF in any trade; while hard working trainees who lack the ability to succeed in one trade are often recommended for reassignment to a different trade.

To the bright but vocationally dissatisfied young trainee, awareness of these informal policies can logically lead him or her to the conclusion that the route to reassignment involves convincing the instructor that, in spite of his or her best efforts, the course material was beyond comprehension. Hence, it is possible that the student may intentionally fail his or her initial basic trades training course, while purposefully avoiding giving the appearance of not being motivated, in order to be recommended for reassignment to another CF trade. Following from this, the course instructor would likely assign such a student to an ability-related failure category, despite the fact the student failed because of a lack of positive motivation to continue in his or her present trade, and not because of a lack of ability.

While plausible, this scenario may seem to represent a rather cynical view of the ethical standards of some of our more able enrolees. Alternative explanations which are equally plausible are available. For example, basic trades training instructional staff may be unable, or perhaps unwilling, to identify the reasons for failure as being motivationally related. This could arise from a sense of unease about passing judgement on the psychological or emotional make-up of their students, or perhaps from an enlightened view that it is often not the trainee's fault that the initial trade assignment decision was a poor one. This view would hold that willingness to accept a less preferred trade to gain entry is evidence of positive motivation towards military service in general, and that this argues for retaining the individual if possible. Cognizant of the impact that a "motivational failure" label may have on the likelihood of retention/reassignment, the instructor may deliberately, and for the best of reasons, misclassify the student as an ability-related failure.

The above discussion points to the need to incorporate a wider array of inputs into the trade assignment decision process. While aptitude is an important element, non-cognitive factors such as vocational interest, temperament, etc., play an important role in determining training success. Rather than treating these factors as "error", and removing them from analyses which are constrained by a narrow, aptitude-based focus, steps need to be taken to standardize and systematize non-cognitive inputs to the process so that matching the person to the job is based on a broader and more complete picture of the relevant factors.

It is suggested that the assumption that maximizing criterion variance will invariably yield better empirical validity estimates tends to be associated with this narrow, aptitude-based focus. The design of this validity study was, in fact, a direct result of acting upon this assumption, since it represents an attempt to maximize the number of records in the data set, while excluding criterion data which were "contaminated" by motivational influences. Improper identification of trainees who should be included in the data set can occur for a number of possible reasons. Whatever the reason underlying this phenomenon, it seems clear that such misclassifications can and do occur.

The measurement of non-cognitive variables which are relevant to the selection/decision-making process is not, in itself, a particularly difficult task. However, this class of measures is much more susceptible to influence by environmental factors and to faking by individual examinees than cognitive ability tests. This, in the same fashion as aptitude tests, renders the incorporation of these measures into the selection process highly problematic. The answer appears to lie in improving the quality of the counselling provided as part of the recruitment process. Organizational recruitment involves two related but separate decision processes, that of the organization, and that of the applicant. Cognitive tests clearly can and should play a major role in the organizational decision about accepting a given applicant. Counselling relates to the individual's decision, and it is here that motivational inputs are most relevant and most easily incorporated. The CF Career Information System (Ellis, 1983), the US Army (JOIN) System, and the US Navy Personnel Accessioning System (NPAS) are all attempts to systematize and incorporate both cognitive and non-cognitive elements into the total accessioning process. Until such systems are fully in place, and reliable data on motivational factors can be collected, the narrow, aptitude-based focus will likely continue to prevail. However, motivational issues will nevertheless continue to impact on training success.

Consequently, it is important to be aware of this potential source of data contamination within the widely accepted practice of attempting to maximize criterion variance when conducting predictive validity studies. Is maximizing criterion variance always best? Yes, but only if those records in the criterion data, that contain information on unsuccessful trainees, are free of misclassification errors.

REFERENCES

Ellis, R.T. (1983, October). Improved recruit counselling with the Canadian Forces Career Information System. Paper presented at the 25th Annual Conference of the Military Testing Association, Gulf Shores, Alabama.

Ellis, R.T., & Saudino, D.A. (1980). Development of a selection model based on the experimental classification battery-men (Working Paper 80-13). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

McMenemy, J.P., Amyot, K.A., & Enkurs, I.A. (1984). Criterion data collection system: Policy and maintenance procedures (Working Paper 84-7). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Miller, D.L., & Angus, R.J. (1984). Predictive validity of selection tests for the Steward 862 trade (Technical Note 20/84). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Miller, D.L., & Angus, R.J. (1985). A revalidation model for the Canadian Forces Classification Battery (Working Paper 85-2). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Literacy, Readability and Knowledge

Frederick R. Chang

Navy Personnel Research and Development Center
San Diego, California 92152

The fields of reading and reading instruction have been highlighted by a number of "great debates." The most popular of the debates was the one made famous by Jeanne Chall (1967) concerning the relative merits of phonics vs. whole-word decoding in reading instruction. Another debate, which has occurred in the military, has concerned the issue of whether or not the armed services should provide basic skills instruction to those personnel who may need it (Sticht, 1982). The papers in this symposium reflect elements of two other distinctions that I feel are important to the issue of remedial education in the Navy. The first issue concerns the distinction between general literacy vs. specific literacy, and the second issue concerns the distinction between reading "skills" and the content knowledge that readers possess. I will comment on these two issues briefly, and will describe some recent research that we have carried out that is germane to these issues.

General Literacy vs. Specific Literacy

The paper by Dr. Moracco discussed the Navy's attempts to move from general literacy training to more specific literacy training, and indeed this is a distinction that is fundamental to this symposium. The paper by Dr. Idar discussed some of the policy implications of this change, and I would like to focus my discussion on some of the cognitive aspects. In teaching "general literacy," one typically refers to the process of giving readers general reading skills that they can then transfer to the reading of materials in many different content domains. The teacher of general literacy is seen as empowering the reader with a set of reading skills that will generalize to the efficient reading of textual materials in any subject. In general the public school system teaches general literacy to its students. In teaching "specific literacy," on the other hand, there is the recognition that people read texts for specific purposes and in specific content areas. The specific literacy teacher attempts to teach the reader what he or she needs to know in order to perform a given task. Since it is generally held that it takes knowledge to get knowledge, it therefore stands to reason that if you give instruction to a reader in a content area, it will be easier for that reader to learn new information themselves in the content area. Most civilian adult basic education programs subscribe to some form of the specific (functional) literacy approach.

The clash between the approaches comes in dealing with the issue of remediation. Basic skills training in the military has had strong supporters on both sides of the issue (see Sticht, 1982 for a review). The detractors from the specific literacy approach claim that in teaching students about specific domains, the students will not acquire the necessary skills to read efficiently in all domains in which they will need proficiency. The fear is that the students will be artificially "propped up" in a particular domain, only to fail, miserably, when asked to read in another domain. The supporters of the specific literacy approach to remediation point out that general literacy training for children in the public schools and general literacy training for adults in the armed services are quite different. Children in the public schools have many years to gain literacy skills; young adults in the military do not. Given that one knows quite well the sorts of literacy tasks that young adults in the military are to perform (i.e., reading for learning or reading for doing something), then it makes sense to provide basic skills instruction in the context of the specific literacy task. After all, in general literacy students are asked to read in many different

content areas, why not have them reading in the domain in which they will be working?

The debate will not soon be settled, but it is interesting to ask the question: What is learned in general literacy? The answer to this question is certainly most complex, and can be addressed at many different levels. At the lexical level, though, it is certainly the case that one goal of general literacy is to provide readers with a large enough vocabulary so that they can read in many domains. Nagy and Anderson (1984) have argued convincingly that it is simply not possible for people to obtain the vocabulary that they have through direct instruction in vocabulary. There are simply too many words; it would take too long. It seems reasonable to suppose then, that much vocabulary is acquired by reading in many different content areas (Nagy, Herman & Anderson, 1985).

How much vocabulary is needed to read in a variety of different content areas? An analysis of the vocabulary used in many different publications reveals that relatively few words are used most commonly. A variety of different word lists have been assembled that contain the majority of the vocabulary in common usage. The different lists are naturally quite similar, and contain relatively few root words (a few thousand). It is generally found, that these words account for the large majority of general use in many popular documents. It was of some interest to me to compare a general literacy text, with a specific literacy text to see what proportion of the words contained in each, appeared on a common word list. That is, it is of interest to know if general literacy texts contain a vastly higher proportion of common vocabulary than specific literacy texts. If so, then one might argue that students in reading specific literacy texts might not be exposed to and learn the common vocabulary needed for efficient reading in other domains. For the general literacy text I chose the comprehension paragraphs from the Gates-MacGinitie reading test level D. The Gates-MacGinitie test is one of the most widely used tests of general literacy in the United States and is the general literacy test administered to all Navy recruits. For the specific literacy text, I chose some passages from the Experimental Functional Skills Program reading assessment battery. The general and specific literacy texts contained roughly the same number of words. For the common word list, I selected the 4908 most frequent words on the Kucera and Francis (1967) list. (These words corresponded to a frequency of occurrence of 20 or greater.)

The results were that 86% of the words on the general literacy text were on the common word list, and 90% of the words on the specific literacy text were on the common word list. This is an interesting finding and suggests that in general most of the words from both texts are ones that are used most commonly. The 14% that are not, from the general text, presumably derive from some particular content area (after all, all the paragraphs on the Gates-MacGinitie test are about some domain). Similarly, the 10% of the words from the specific text that were not on the common word list are presumably specialized Navy words. An implication from this very informal finding is that, at least at the lexical level, perhaps much of what is thought to be learned only in general literacy is also learned in specific literacy.

Processing Skills vs. Knowledge

Another distinction, closely related to the general versus specific literacy distinction, concerns processing skills used in reading versus the knowledge that the reader has of the domain to be read. Research in experimental psychology has been dominated by a concern for the processing skills used in reading. The concern has been both methodological (e.g., Chang, 1983; Kieras and Just, 1984) and substantive (e.g., Crowder, 1984; Gibson and Levin, 1975). One outcome of the emphasis on processing skills has been that some of the findings from the basic research laboratories have been interpreted fairly literally into training approaches with sometimes negative results. One example of this outcome concerns the finding that good readers, in general, exhibit very different eye movement patterns from poor readers. Good readers: 1) make fewer eye movements per line, 2) have shorter fixation durations, and 3) make fewer regressive eye movements, than poor readers. On the basis of these findings, several eye movement training programs were developed. The thinking was that all that one needed to do was to train the poor readers to move their eyes in reading just like good readers, and the poor readers would then be reading just

like the good readers. Naturally, such training failed miserably (see Gibson and Levin, 1975).

This example is not meant as an indictment against the study of basic reading processes -- clearly such study is important. However, the emphasis on the study of reading processes has lead in part to the development of some "process oriented" remedial reading programs. For many readers, these programs have not been successful. The more general lesson to be learned from this is that process differences between groups of readers do not necessarily translate well into instructional programs. Such process differences may be only symptomatic of more profound differences.

In recent years, there has been an increasing emphasis in cognitive science research in the study of the importance of knowledge in cognition. More and more studies in cognitive psychology treat prior knowledge not as a nuisance variable to be eliminated in the study, but rather as the important variable to study. In part, the recent emphasis on knowledge based research has been spurred on by developments in artificial intelligence research. In researchers attempts to make computers smart, they have found that knowledge of the domain plays a critical role (Barr and Feigenbaum, 1984). This agrees quite nicely with the finding that expert and novice human chess players do not differ in general processing skills (i.e., short term memory processes), instead they differ in that expert chess players have a great deal of highly organized domain specific knowledge about chess positions. Increasingly, researchers in reading and reading instruction are recognizing the importance of studying the prior knowledge of the reader, (e.g., Chiesi, Spilich and Voss, 1979; Kieras and Johnson, 1984) and recognizing the importance of the tradeoff between processing skills on the one hand and knowledge on the other and what the implications are of this tradeoff to remedial training.

Knowledge and Readability

The concepts behind the functional reading assessment battery, described in the paper by Dr. Sticht, have been extended for use in estimating the reading demands of documents. I will briefly describe some exploratory research that we have done (this work was done in collaboration with Dr. Sticht), as it begins to shed light on the role of prior knowledge on reading and readability.

Readability researchers have for many years been interested in determining ways to best match readers to documents. That is, the goal has been to provide a way to select a document for a reader so that the document is not too difficult (or too easy) for the reader to comprehend. The result of the research has been among other things, the development of many different readability formulas (see Klare, 1974 for a review). Very briefly, the formulas attempt to relate surface structure characteristics (e.g., sentence length and word length) of the document to comprehensibility. It has been found empirically that these characteristics are correlated with measures of comprehensibility (cloze tests, multiple choice tests etc.). While there have been many criticisms of these formulas (e.g., Duffy, 1985) they continue to be used.

One major problem with these traditional formulas is that they treat all readers identically. That is, a given document is presumed to be equally comprehensible to all readers regardless of a reader's prior knowledge of the subject matter. Clearly this is a profound problem. Therefore, the goal of our research was to assess the influence of prior knowledge on the comprehensibility of a document and to develop new readability formulas that take account of this prior knowledge.

We studied a sample of 296 Navy recruits who had taken the Navy knowledge test as part of the functional reading assessment battery (described in the paper by Dr. Sticht). This test does not involve reading any paragraphs; subjects are simply asked questions about the Navy, thus the test measures their prior Navy knowledge. Based on the results of this test subjects were broken into four groups (of roughly equal size) that reflected performance on the test. Group 1's (low knowledge) performance was lowest on the knowledge test and Group 4 (high knowledge) was the best. Groups 2 and 3 were intermediate.

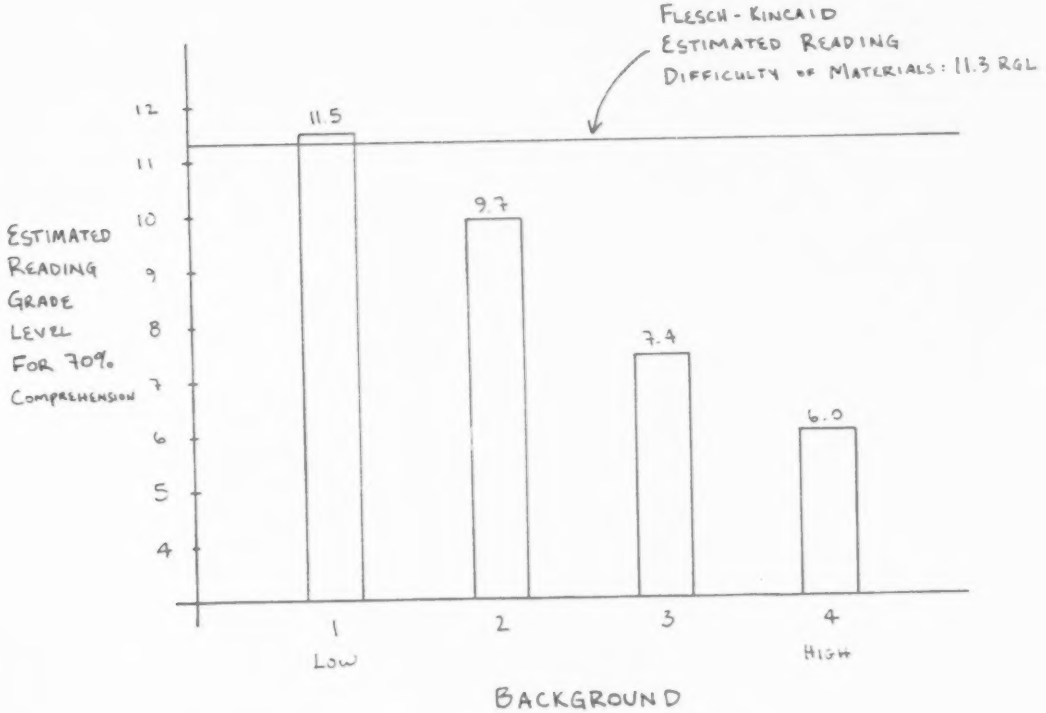
All groups read eleven different Navy related passages. For three of the passages subjects performed a reading-to-learn task (the multiple-choice questions had to be answered from memory; the passages were not available for question answering), and for the other eight passages

the subjects performed a reading-to-do task (the passages were available for consultation in answering the comprehension questions). In the preliminary results reported here, all eleven passages were collapsed into a single analysis; in a later report we will break these apart.

We performed a standard readability formula analysis (see e.g., Caylor, Sticht, Fox & Ford, 1973) but we did it separately for each of the four knowledge groups. For each knowledge group we determined the reading grade level (RGL) required to comprehend each passage at the 70% level (a linear interpolation method was used to determine scaled RGL). Then four different surface structure features of the passages were used to predict scaled RGL: average sentence length, average word length, proportion of content words and average content word length. The four different formulas differed primarily in their intercept and were combined into single formula that contained surface structure components and a knowledge component.

The results are clear in showing three general trends: 1) in general, the more readers knew about the domain, the easier the passages were to comprehend, 2) traditional readability formulas will overestimate the readability demands of a passage for those who have some prior knowledge in the content area, and 3) prior knowledge accounted for more of the variance in the prediction of comprehensibility than did the surface structure features.

Using the new knowledge-based readability formula the estimated reading difficulty of each passage was computed for each group (the estimates are in RGL units). The estimated difficulty averaged over all passages is presented in Figure 1 for the four knowledge groups. That is, for each knowledge group the figure shows the estimated RGL needed to comprehend the passages at the 70% level. The solid line at the top of the figure shows the estimated reading difficulty of the same passages by the standard DoD Flesch-Kincaid readability formula (Kincaid, Fishburne, Rogers & Chissom, 1975).



Consider the low knowledge group. One can see that the estimated reading difficulty for the low knowledge group is almost identical to the reading difficulty estimated by the Flesch-Kincaid formula. This indicates that in considering readers with very little prior knowledge of the content area, the standard DoD readability formula predicts actual performance relatively well. However, as prior knowledge of the content increases one can see that the traditional formula consistently overestimates the RGL needed to comprehend the passage. Thus, the traditional formula treats all readers as if they have little or no knowledge of the domain. If the reader does know something about the domain, then a given document will be easier to comprehend than would be indicated in a traditional formula.

In computing the knowledge-based readability formula we found that the surface structure variables accounted for roughly 26% of the variance in predicting the scaled RGL values. The knowledge variable, on the other hand accounted for about 43% of the variance. Thus, the knowledge variable accounted for a substantially greater proportion of the variance, and this difference was statistically reliable.

The claim has been made in this research that the more readers know about the domain to be read, the easier a passage is to comprehend. It should be pointed out that readers in this study evidently did indeed read and extract information from the test passages and did not simply answer the comprehension questions without reading the passages. An indication of this comes from the fact that overall performance on the Navy knowledge test (no reading passages involved) collapsed over all groups was 47% while test performance based on the reading passages was 74%. Thus, the information extracted and learned from the passages accounted for a sizable increase in test performance.

It should be emphasized that these are only preliminary results, and that much work needs to be done. However, these findings are potentially extremely important. We have demonstrated that prior knowledge of a content area has a large effect on the comprehensibility of a particular passage, and we have been able to quantify that effect in a precise way. We have also developed a technique that can be used quite generally to assess the role of prior knowledge in any domain and incorporate the knowledge influence into a readability formula. The new technique will allow us to construct a variety of different formulas, based on different measures of prior knowledge, to assess the readability of both technical and instructional documents. In this way we will be able to match better, readers to documents taking into account what the readers already know about what is to be read.

REFERENCES

Barr, A. & Feigenbaum, E.A. (1981). *The handbook of artificial intelligence, Vol. 1.* Stanford, CA: HeurisTech Press.

Caylor, J.S., Sticht, T.G., Fox, L.C. & Ford, J.P (March 1973). *Methodologies for determining reading requirements of military occupational specialties.* (HumRRO Tech. Rep. 73-5). Alexandria, VA: Human Resources Research Organization.

Chall, J. (1967). *Learning to read: The great debate.* New York: McGraw-Hill.

Chang, F.R. (1983). Mental processes in reading: A methodological review. *Reading Research Quarterly, 18,* 216-230.

Chiesi, H.L, Spilich, G.J. & Voss, J.F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 257-273.

Crowder, R.G. (1982). *The psychology of reading: An introduction.* New York: Oxford University Press.

Duffy, T.M. (1985). Readability formulas: What is the use? In T. Duffy & R. Waller (Eds.). *Designing usable texts.* New York: Academic Press.

Gibson, E.J. & Levin. H. (1975). *The psychology of reading.* Cambridge, MA: The MIT Press.

Johnson, W. & Kieras, D.E. (1983). Representation-saving effects of prior knowledge in memory of simple technical prose, *Memory & Cognition, 11,* 456-466.

Kieras, D.E. & Just, M.A. (Eds.) (1984). *New methods in reading comprehension research.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L. & Chissom, B.S. (February 1975). *Derivation of new readability formulas (Automated Readability Index, Fog count, and Flesch Reading Ease Formula) for Navy enlisted personnel.* (CNTT Research Branch Rep. 8-75), Millington, TN: Chief of Naval Technical Training.

Klare, G.R. (1974-1975). Assessing readability, *Reading Research Quarterly, 10,* 62-102.

Kucera, H. & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press.

Nagy, W.E. & Anderson, R.C. (1984). How many word are there in printed school English? *Reading Research Quarterly, 19,* 304-330.

Nagy, W.E., Herman, P.A. & Anderson, R.C. (1985). Learning words from context. Reading Research Quarterly, 20, 233-253.

Sticht, T.G. (June 1982). *Basic skills in defense.* (HumRRO-PP-3-82). Alexandria, VA: Human Resources Research Organization.

Footnotes

The views, opinions, and/or findings contained in this paper are those of the author and should not be construed as an official Department of the Navy or Department of Defense position, policy or decision, unless so designated by other official documentation.

The work on readability was done in collaboration with Dr. Thomas Sticht; a more formal, co-authored paper describing this research is forthcoming. I am grateful to Dr. William Montague for generally useful discussions concerning the role of knowledge in cognition and to Dr. Edwin Aiken for helpful comments on an earlier draft of this paper. Thanks are due Tami Lopez for assistance in the data analysis.

The XFSP: Reading and Mathematics Project

Thomas Sticht, Louis Armijo, Natalie Koffman, Kent Roberson
U.S. Naval Postgraduate School

In an ongoing program of instructional development, the authors and other colleagues from the Naval Postgraduate School and the Navy Personnel Research and Development Center are developing 45 hour developmental reading and mathematics programs for the Navy under sponsorship of the Chief of Naval Education and Training. The programs we are developing will replace programs being offered by some dozen contracting organizations world-wide. The contractor programs are "general" in their orientation, whereas the programs we are developing are Navy-related, and they are oriented to assist enlisted personnel meet requirements for career promotions to higher paygrades and responsibilities.

In this paper, which will focus on the XFSP (Experimental Functional Skills Program): Reading, I will first discuss the conceptual framework for the development process, including a simple model of the *human cognitive system* and the concept of *functional context training*. These conceptual frameworks help to give the development process broad, general directions. In this regard, they provide what is typically missing in current instructional systems design (ISD) methods, and that is a view of human cognition and the conditions for its use that can provide heuristics for needs assessment, task analysis, and program design.

Following the discussion of the underlying concepts for program development, I will briefly describe the program we have developed as it stands at the present time, and then I will describe the assessment test battery we have developed to assess program effectiveness. Finally, a small scale study to compare the Navy-related reading program to the general reading program offered by a local contractor is described.

Human Cognitive System Model

In conducting the development of the XFSP: Reading program we have worked from a stripped-down, simplified model of a human cognitive system and the processes the system uses for extracting and representing information in the environment. The model is schematized in Figure 1. The model contains three major components, two of which are "inside the head", and the third component which is "outside the head" and includes various information displays produced by or comprehended by the cognitive components. The latter include the ***knowledge base*** , which is a long term memory that contains all the information and knowledge possessed by the person, and the ***processing skills*** , including language, that operate on the information in both the knowledge base and the external environment to produce comprehension, communication, and thinking. The model as depicted in Figure 1 is of one who possesses both oracy and literacy information processing skills. The literacy processes include all those used to recode written language into internal forms comparable to those used in oral language, and, in addition, to perform all those literacy tasks that are not instances of writing as a second signaling system for speech. The tasks unique to literacy are those made possible by the properties of graphic displays: they are more or less permanent (thereby permitting study), they can be arrayed in space (permitting the construction of forms, signs, flow charts, graphs, etc.), and they can use the properties of light (contrast, color) to guide attention and facilitate information processing.

According to the model of Figure 1, then, the performance of literacy tasks requires knowledge about what one is reading or writing (including mathematics knowledge when reading in that domain); processing skills for thinking about what to communicate or for comprehending what others communicate; and, of course, graphic displays of information in

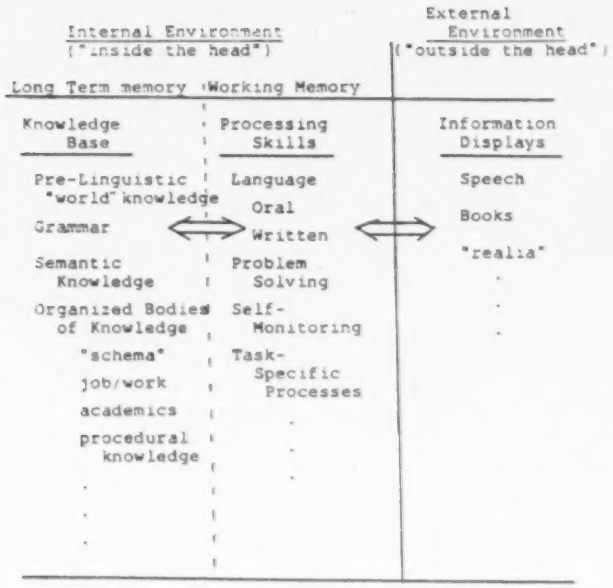


Figure 1

Cognitive system model for instructional program development.

the environment to be processed for meaning. An important implication of this analysis is that it reveals that literacy, considered as the ability to comprehend and use the graphic symbol systems of writing, graphing, illustrating, mathematics, and so forth, can be enhanced by improving either one's knowledge base in a given task domain, or one's processing skills,or, as in the case of improving the readability of materials, through the redesign of the graphic information displays the cognitive system must deal with, or a combination of these factors. Use of this conceptual framework in determining the learner's needs in literacy programs is illustrated later on, following a discussion of the *functional context* approach to instructional design that has undergirded the XFSP:Read program development activities.

The "Functional Context" Concept

Along with the simple cognitive model of Figure 1, we have followed what is known by some as the "functional context" approach to education and training development (Shoemaker, 1967). The essence of this approach is contained in two major goals for instruction. First, always try to make the instruction as meaningful to the learner as possible in terms of the learner's prior knowledge. This facilitates the learning of new information by making it possible for the learner to relate it to knowledge already possessed, or to make it possible for the learner to transform old knowledge into new knowledge. Second, as much as possible, use the materials and equipments that the learner will use after training or education as part of the instructional program. This will motivate the learner by showing that what is being learned is relevant to a future goal, and it will promote transfer of learning from the classroom to the next training or "real world" activity. In short, the functional context method of instructional design attempts to motivate and promote learning and transfer by making the program meaningful in terms of the learners past, present, and future.

The XFSP: Read Program

The XFSP: Read program was developed following the guidance of the concepts described above applied in studies of what kinds of tasks Navy personnel perform using reading skills in training and job settings. In this research, students, instructors, and job performers in ten Navy jobs were interviewed and asked for information regarding two major types of reading tasks: reading-to-do something and reading-to-learn something. In a reading-to-do task, the person is performing some job task, needs some information from a document, looks-up the information, holds it in working memory long enough to apply it, and can then forget it. In a reading-to-learn task, the person reads information to be stored in long-term memory as part of the knowledge base, and then retrieves it (or a reconstruction of it) for use at some later time, such as taking an end of week test, or for performing a task on the job.

The interviews with personnel revealed that reading-to-do and reading-to-learn were performed to about the same extent in school situations, but on the job reading-to-do comprised about three-fourths of the reading tasks. It was also found that the processing skills performed in reading-to-learn were more complex than those used in reading-to-do. Whereas the latter emphasized information location and extraction skills, such as use of tables of content, indexes, "thumbing" or "flipping" through or searching tables and figures, reading-to-learn involved more elaborate activities to merge new information with old knowledge. The primary reading-to-learn processes were categorized into four groups: (1) reread or rehearse processes, in which learning was accomplished by rereading some portion of the material, or was repeated in some way over again to oneself; (2) question/problem solve processes, in which learners asked themselves questions about the material, or solved problems in textbooks; (3) relate/associate processes, in which the learners transformed what they read into some other form, either by paraphrasing, making internal images, watching a movie or demonstration and relating what they had read to the new information gained from these "iconic" or "realia" information displays (see Figure 1); and (4) focus attention processes, such as highlighting with colored pens, underlining, summarizing or some similar methods for focusing attention on a limited aspect of the material, usually in conjunction with a reread or rehearse activity later on.

The interviews also revealed the role of the knowledge base in performance of reading tasks. For instance, it was found that close to 60% of job tasks involving reading had been performed previously, and for about half of the 325 reading tasks cited by the sample of 178 personnel, additional reading related to the task had been performed, and for two-thirds of these cases, the related reading helped in reading the material cited in the interview (see Sticht, 1977 for a more complete description of this work).

The Career Progression Reading Program

On the basis of the foregoing research and additional study of the reading demands of the Navy environment, we have designed and developed a reading program that has a functional context for Navy personnel in that the program uses Navy content derived from materials they must study to pass promotion tests, and the information processing skills are of immediate use to them. Most of the students are "mid-level" literates, with reading skills in the 6th to 10th grade levels, and most have had one one or more years of duty, and so they are able to relate their knowledge base to the content of the course.

The instructional program is designed to be delivered either by a teacher using paper-and-pencil materials in a classroom, or by a teacher, paper-and- pencil materials and computers

in a classroom, or in a learning center using computer-based instruction alone. Figure 2 shows the teacher-, book-, and computer-based instructional classroom delivery system that is the central model for the delivery system. The basic course is three weeks in duration and students attend for three hours a day for a total of 45 hours of instruction (the time for instruction was specified by the Navy Campus office). The books include a special reader, which, based on the model of Figure 1, we call a Navy Knowledge Base. This reader contains extracts and revisions from the technical manuals that personnel must read and learn from to pass promotion tests or perform higher level duties. The contents were selected because they contain the information that is on the practice tests in the manuals and are deemed important by the Navy management.

The second book is called the Information Processing Skills book, and it is modeled after the processing skills component of the model cognitive system. This book presents lessons and practice in performing various processing skills for reading-to-do and reading-to-learn tasks using the Navy Knowledge Base book. The goal here is to present *externally* a knowledge base and processing skills to serve as didactic tools for talking about the *internal* knowledge base and processing skills and to hence make students aware of their cognitive systems and how to apply them to doing and learning literacy tasks.

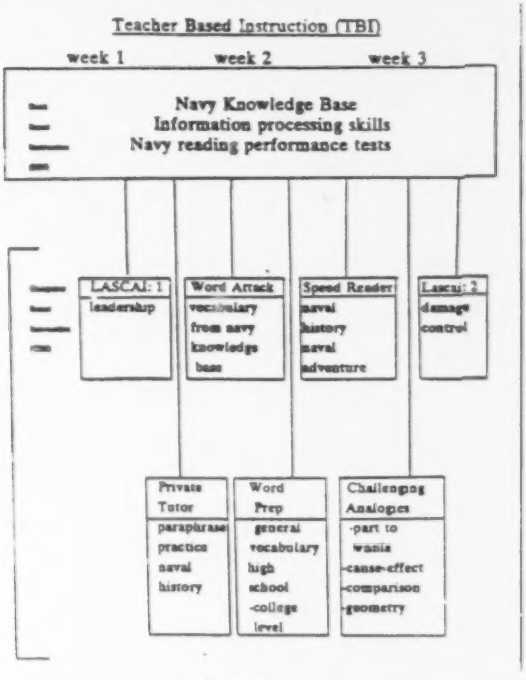


Figure 2

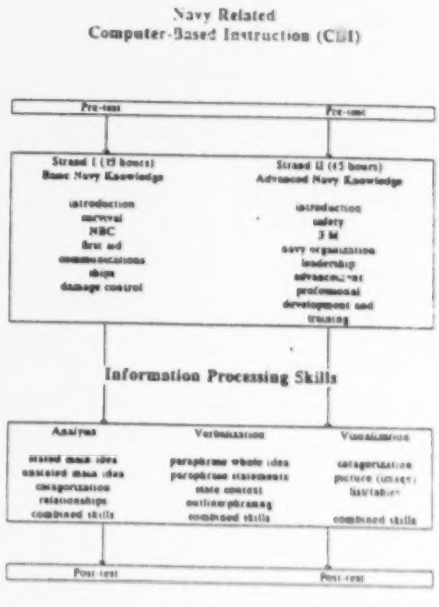


Figure 3

The computer based instruction consists of public and private domain programs that serve as sort of "electronic worksheets" that the teacher can assign students to do while he or she works with other students on a small group basis. The computer software was selected to permit the teacher or curriculum specialists to personalize the instruction, and so each program has an editor feature that permits one to enter content works or paragraphs for the specific literacy domain they wish to teach. The Lascai, Word Attack, Speed Reader, Word Prep, and Ghallenging Analogies programs all have a game format that operates automatically on the content that is edited into the program. The Private Tutor is sold by IBM and is a very inexpensive, easy to use authoring system for preparing indivdiual lessons or entire computer-based instructional courses.

Figure 3 shows the computer-based instruction that has been developed for use in a learning center as a stand-alone, self-paced instructional program. The program has two Strands, each providing about 15-20 hours of instruction depending upon the reading skill level of the user. The Strand I material contains content appropriate for sailors seeking promotion to lower level supervisory positions, while the Strand II material is primarily for those seeking higher level promotions (these are only approximations, since some contents are useful across

the board in both strands). The bottom part of Figure 3 shows the processing skills taught in both strands of the computer-based instruction.

Functional Reading Assessment Battery

In addition to the funtional, Navy-related, standardized teacher-, book-, and computer-based instructional programs, we have also developed a new reading test battery to test Navy-related reading skills (reading-to-do and reading-to-learn) as well as Navy knowledge. The latter, that is, knowledge gained, is rarely measured in reading programs because reading is considered as a content-free, process skill. But the cognitive model of Figure 1 makes clear that knowledge of what one is reading is required to make reading comprehension possible. So we are assessing the improvement in knowledge as a function of participation in the functional reading program. Figure 4 summarizes features of "traditional" reading tests and our new "functional" reading battery.

An Evaluation Study

In a small evaluation study, the improvement of a sample of students who took a "general" reading program offered by education contractors was compared to the improvement of students in our "functional" reading program on three tests: a general reading test that gives grade levels of performance, our Navy reading-to-do test and our Navy Knowledge test (the reading-to-learn test was not available at this time). The results are summarized in Figure 5. The data show that, in general, people tend to learn what they are taught. The "general" reading program did better on the general reading test, but this did not transfer to the Navy reading and knowledge tests to any significant degree. The Navy-related reading program, on the other hand, resulted in only a little improvement on the general reading test, but made consistent improvements where it counts for Navy personnel, that is, in their Navy reading and knowledge.

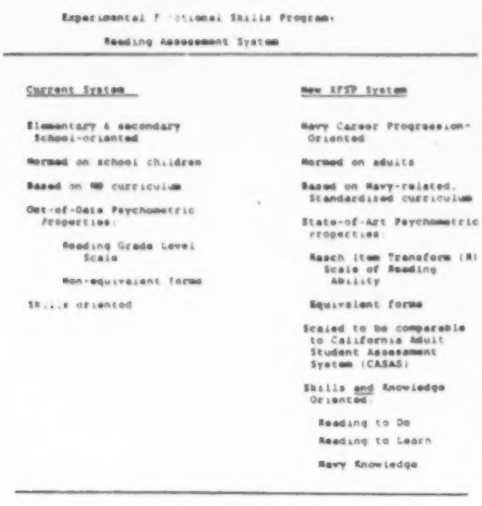


Figure 4

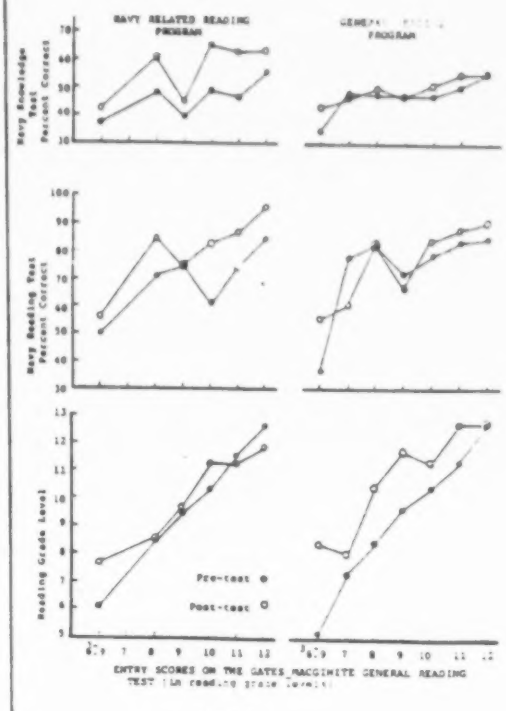

NAVY RELATED READING PROGRAM
Navy Knowledge Test Percent Correct
Navy Reading Test Percent Correct
Reading Grade Level
Pre-test
Post-test
ENTRY SCORES ON THE GATES MACGINITE GENERAL READING TEST (in reading grade levels)

Figure 5

Future Tasks

At the present time we are polishing up the XFSP:Read materials and packaging them for implementation and further evaluation. We have also started the development of an XFSP:Mathematics program based on the same general conceptual framework discussed herein. Finally, we have initiated some exploratory study of the use of the new Navy reading assessment battery in developing an approach to estimating the reading demands of Navy training and jobs using the reading-to-do and reading-to-learn tasks as separate criterion variables, and including Navy knowledge and other knowledge domains as variables in the predictor along with the more traditional passage-based variables (e.g., word and sentence length) used in readability formulas.

Reference

Shoemaker, H. (1967). The Functional Context Method of Instruction. HumRRO Professional Paper No. 35-67, Alexandria, VA: Human Resources Research Organization.

Footnote Colleagues who have worked on one or more aspects of this project include Drs. R. Weitzman; Frederick Chang; William Montague; and Monique Allen; Joan Molleur; Susan Wood; Jim Joelson; Jim Powers; Tom Kasten; and Otto Kruse. Project Monitor is Dr. Judy Moracco of CNET/TTI.

A CONCEPTUAL FRAMEWORK FOR OCCUPATIONAL EXPLORATION

John J. Pass, PhD

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Introduction

The need for improved dissemination of military career information for purposes of occupational exploration and person-job matching (PJM) has long been apparent. With the advent of voluntary forces, the military was removed from automatic consideration in the career plans of the nation's youth. The problem has been to provide young people with the means of career exploration in order to acquaint them with the benefits of military training and career opportunities.

Basic Components

Essentially, an occupational exploration system must include: (1) personal assessment procedures to measure the individual along a number of cognitive and noncognitive dimensions in order to provide selection information to the organization and enhance the individual's self-awareness; (2) occupational information that is accurate, timely, sufficiently comprehensive, and personalized to the individual; and (3) a bridging mechanism between personal information and job information that might include a vocational guidance or counseling process.

The final design of any occupational exploration system will be determined by its purposes, the constraints of cost and technology, and the ingenuity of the designer. Figure 1 illustrates some examples of the types of information that could be included. First, there are the individual dimensions or attributes to be measured. Second, there is the occupational information that would be useful for an individual to use in gaining some understanding of requirements and conditions of occupations or careers. Third, typical search strategies that are useful in the attempt to match the individual characteristics with organizational and occupational requirements are shown.

Search strategies are a critical component of an occupational information system, one that deserves careful consideration. Upon the search strategy rest many of the system's time and cost requirements. Equally as important is the effect of the search strategy on the person-job match. Any effective procedure for matching persons with jobs must take into account both individual and institutional characteristics. Individual characteristics include abilities, preferences, interests, and goals. Institutional characteristics include priorities, objectives,

BEST COPY AVAILABLE

training program vacancies, and personnel requirements. An effective search strategy, then, enables a narrowing down of alternatives so that information is developed or presented for those occupations that are best suited to the individual. Ultimately, the search strategy conduces to a tentative choice of an occupation or an occupational field.

System evaluation is an additional component of an occupational exploration system. This can prove to be an important element. One part of this component can be a program that seeks to obtain user evaluation of the system. A series of multiple-choice questions may be administered on-line to assess general satisfaction with the guidance process and the interactively programmed computer system. Other, more formal evaluative research can also be employed. The information from both sources can be used as feedback to help improve the system.

APPLICANT ASSESSMENT

-Aptitude
-Interests
-Medical Evaluation
-Work Experience
-Career Maturity

OCCUPATIONAL INFORMATION

-Job Requirements
-Training Provided
-Physical Demands
-Career Benefits
-Job Openings

SEARCH STRATEGIES

-Accessing Occupational Information
-Matching Personal and Occupational Characteristics
-Narrowing of Options

SYSTEM EVALUATION

-User On-Line Evaluation
-Interview
-Impact Evaluation

Figure 1. Occupational Exploration System

Current Situation and Prospects

Unfortunately, although the need has been recognized for many years, occupational information is frequently scanty or outdated, and difficult to access. Many times, there is a lack of individual assessment methods so that personal considerations can be dealt with. Furthermore, current occupational exploration

systems in both civilian and military settings are often notably deficient in provision of accurate and comprehensive military occupational information.

However, there are a number of bright spots in this setting. One of them is computerization. Computerization offers a major hope for amelioration of the deficiencies in availability and accessibility of accurate, timely, and personalized occupational information. Computers can provide immediately scored self-assessment instruments, didactic counseling materials, and comprehensive, up to date, rapidly retrieved, personalized job information. Computers can deliver standardized occupational information to the job applicant, thus minimizing or at least limiting the introduction of information bias. Individuals benefit from complete information and impartial guidance in selecting an occupation, and counselors are freed of the need for broad and current knowledge of military occupations and training.

Furthermore, in the civilian community, the past generation has witnessed the development of excellent computerized occupational information systems. And, more recently, reliable military career information has been made available under the sponsorship of The Office of The Assistant Secretary of Defense (FM&P), in support of a policy advocating improved guidance and counseling, and the provision of honest, accurate military career information to high school and college students as well as to prospective enlistees.

The Military Setting

Within the military research community, there is progress in the practical evolution of systems and techniques to accomplish those ends, with the design and development of systems and techniques that have wide applicability in a variety of civilian and military settings. Let me cite a few examples that incorporate various components shown in Figure 2 above.

In FY75, the Navy Personnel Research and Development Center (NAVPERSRANDCEN) began work on an advanced developmental effort called Project CONTRACT (Computerized Navy Techniques for Recruiting, Assignment, Counseling, and Testing). Two products of CONTRACT were the development and demonstration of a computerized system called the Navy Vocational Information System (NVIS) and the development of an optimal personnel assignment algorithm for use with the computerized entry level job assignment system employed by the Navy.

One research objective of NVIS was to develop a prototype interactive computerized occupational information system that would provide young men and women with personalized occupational guidance and a list of related civilian and Navy jobs that demonstrated a good match between their own attributes and typical job requirements.

NVIS databases contained information on 279 civilian jobs, 114 worker trait groups, 79 entry level Navy occupations, and more than 100 specialized Navy jobs. NVIS maintained an interactive dialogue with the user via a cathode ray tube.

NVIS proved to be the precursor to an expanded Navy occupational information system because an outgrowth was the conceptualization and advanced development of a microcomputer based person-job matching system, another prototype computerized vocational guidance system called AGENA (Automated Guidance for Enlisted Navy Applicants).

AGENA used interactive programming to lead the person through a logical, thought-provoking dialogue that introduced the system and equipment; proceeded through preliminary aptitude screening; progressed, via learning how to plan for a career and discover personal interests and aptitudes; and ultimately explored a number of Navy entry-level occupations that matched personal interest and aptitudes, with the opportunity to assess the availability of options in a rather wide timeframe.

Within the broad function of person-job matching, three subfunctions were supported in the AGENA system: (1) aptitude screening; (2) vocational guidance; and, (3) assignment prediction. AGENA took personal and organizational factors into consideration and, coupled with information on when the individual wanted to enter the Navy, determined which entry level jobs were most appropriate for meeting individual and organizational considerations. The availability of these entry-level assignments was indicated for a three-month timeframe.

AGENA information on the entry-level Navy jobs was available in two formats. The abbreviated version, designed for video terminal display, included five sections: (1) general description, (2) related civilian jobs, (3) qualifications, (4) working conditions, and (5) Navy opportunities. An extended description, available in hardcopy as an option, included all sections of the abbreviated description plus three additional sections: (1) what the people in the rating do, (2) sea/shore rotation, and (3) the training provided by the Navy.

The substantial contributions Navy training can make to total career development discussed in AGENA, along with a brief discussion of the general value of Navy training and experience. On-line access to Civilian Occupations Database gave descriptions of civilian occupations (or clusters of occupations) related to the Navy assignment opportunities on the video display terminal and a hardcopy output from the printer. Descriptions of civilian occupation included: (1) general description, (2) qualifications and training (3) pay wan working conditions, (4) employment outlook, and (5) related Navy jobs. Finally, a brief discussion of additional benefits of Navy enlistment (e.g.,medical benefits) was included.

AGENA was shown in a demonstration version in the fall of 1981. Subsequently, NAVPERSRANDCEN undertook supportive R&D for the Army Joint Optical Information Network (JOIN). The purpose of this research was to design, develop, test, and incorporated in the JOIN system, a computerized vocational guidance system, and to develop other capabilities including computerized adaptive screening and assignment predition. Products of that effort have been incorporated intc JOIN or are under further development.

To the list of military laboratory efforts in the area of occupational information systems would have to be added the important work being carried on by the Canadian Forces Personnel Applied Research Unit, in Toronto. Their Canadian Forces Information System offers several excellent examples of the use of the components outlined above, including orientation video and realistic job preview.

However, the examples I have given were to illustrate the incorporation of the various elements or components into a system that delivers accurate and consistent military occupational information in a user friendly way.

Conclusion

There is growing interest and progress in military occupational exploration. While many issues remain in the development of such a system (e.g., level of occupational information specificity, degree of automation., and the location of the system within the organization), these issues tend to be more easily resolved once the objectives for the system are established. We can certainly envision two clear objectives: to increase the awareness of military career options; and to conduce to more optimal person-job matches that will enhance job satisfaction and productivitry while decreasing attrition and personnel turbulence.

The potential payoff of pursuing this R&D area in terms of increasing the numbers of military applicants and decreasing attrition definitely warrants additional effort. Work is moving ahead on the conceptual design of a comprehensive computerized vocational guidance system that assists the job seeker in self exploration, provides accurate and consistent occupational information, and helps in relating personal characteristics and occupational choices. To enhance awareness of military career options, and make military job training a viable element of career planning to be seriously considered by those entering the work force, is the goal of these projects.

Development of the Career Maturity Assessment

Esther E. Diamond, Ph.D.
Educational and Psychological Consultant
Evanston, Illinois

Introduction

Many recruits entering military service have not engaged in thoughtful career planning, or adequately assessed themselves, or sought, located, and assembled career information, or developed an appropriate set of vocational coping behaviors. Consequently they are not yet ready to make informed occupational choices.

To assist recruits in the career choice process, the design, development, and pilot testing of a career maturity assessment (CMA) instrument was undertaken, for eventual incorporation into a Computerized Vocational Guidance (CVG) system for use in recruiting. CMA scores would determine the point in the guidance sequence to which applicants would branch.

Background

As Crites (1973) has pointed out, prevailing views of vocational behavior before the 1950s were almost entirely nondevelopmental. Vocational decision-making was considered to be a time-bound, largely static, one-time event, which usually occurred upon high school graduation, "when an adolescent took stock of himself and the world of work and then decided what he was going to do" (p. 5). Super's seminal work in vocational development theory (1953,1957) was among the first to introduce the concept of career maturity, defined as the readiness of an individual to make career decisions expected at a particular age. He defined the term further as "...the repertoire of coping behavior leading to outcomes...a developmental rather than an outcome construct" (1974, p.11). That decision-making plays a major role in vocational development has been underscored in the work of most major career development theorists (e.g., Super, 1984; Ginzberg, 1984; Crites, 1973, Osipow, 1983). In general, self-concept or self-knowledge and decision-making skills appear to be common elements in most of the research reviewed. In developing the CMA, then, these concepts were the major focus.

Instrument Development

The Navy Personnel Research and Development Center (NPRDC) specified that CMA items should be written at the sixth-grade reading level and should not exceed 30, with maximum administration time of five minutes; face validity should be high; responses should be true/false or Likert-type; scoring should permit dichotomous reporting of results; scale reliability should be at least .65; and the instrument should be validated against a suitable commercially available instrument.

Following review of the related literature, a pilot version was constructed, with items drawn from a number of sources, including the literature review, the author's professional experience in developing a number of career guidance programs, and the objectives and results of the

National Assessment of Career and Occupational Development (National Assessment of Educational Progress, 1976). The CMA at first consisted of 38 items, to permit later elimination of any that proved unsatisfactory on the basis of item statistics.

Three responses were possible for each statement: Y (Yes, the statement is generally true), N (No, the statement is generally not true), and ? (Uncertain). The Y and N responses had a score value of 0 or 2, depending on the particular item, while question marks had a score value of 1. Following are three examples of the items:

I envy people who don't have to work.

I am afraid to try new things.

I have a good idea of my interests and abilities and how they relate to different occupations.

Variables

The total score on My Vocational Situation (MVS) (Holland, Daiger, & Power, 1980), a 26-item vocational decision-making instrument with three scales, was selected as the criterion variable.

To study the validity of the CMA for subgroups, classification variables were collected for age (under 25, 25 and over), gender (male, female), and level of education (less than high school graduation, high school graduation, some college, four-year college graduation).

Sample

The sample consisted of U.S. Army recruits from nine forts across the country. The total number tested was 405 (323 males and 62 females), but because of errors in the administration, not all recruits completed both the CMA and the MVS. CMA answer sheets for 401 recruits and MVS answer sheets for 304 recruits were analyzed.

Data Analysis

The 401 CMA cases had either no missing responses or less than 25% missing. Missing responses were coded as ? and scored as a 1. For purposes of calculating individual item contributions to reliability, only cases with no missing responses (348) were used. For the MVS, 304 cases were used. Intercorrelations were obtained between all variables. Alpha was computed for total CMA and MVS and for CMA with each item, in turn, deleted.

Results

Correlation between total CMA score and total MVS score was .75, indicating a strong relationship between the CMA and the criterion measure, with more than 56% of the variance (.75 squared) accounted for. Alpha was .83. To reduce the total number of items in the CMA, the individual item data described previously and the correlations with total score were examined. The four items with the lowest correlations with total score and whose

elimination would increase alpha the most were dropped. NPRDC agreed to let the instrument stand at 34 instead of 30 items. A second data analysis was then run.

Alpha for the revised CMA increased to .85, and correlation with total MVS increased to .78, indicating an even stronger relationship than before, with 60 percent of the variance (the squared correlation of .78, which is equal to .6084), accounted for. Total score mean was 51.04, the median 53.02, and the mode 54. The cutoff was set at the median, 53.

The mean for females was higher than that for males, and there was very little difference between means for the two age groups, although the older group scored higher. Means increased with educational level.

Summary statistics for the total CMA sample, based on the 34 items, are presented in Table 1. Table 2 gives means and standard deviations by gender, age, and educational level. The z-ratios computed for the differences between uncorrelated means showed no significant differences between males and females, age groups, or educational levels. Table 3 gives intercorrelations with MVS, gender, age, and educational level.

Table 1

Summary Data for Total Career Maturity Assessment Sample
(N=401; highest possible score = 68)

Mean	51.04	SD	10.73
Median	53.02	SEmeas	.54
Mode	54.00	Skewness	- .75

Table 2

Career Maturity Assessment Means and Standard Deviations by Gender, Age, and Level of Education

Variable	N	Mean	SD
	(By Gender)		
Male	319	50.51	10.97
Female	82	53.08	9.52
	(By Age)		
Under 25	364	50.56	10.62
25 and over	37	55.75	10.76
	(By Level of Education)		
Less than HS grad	57	48.68	11.68
HS grad	205	50.33	10.82
Some college	125	52.41	10.00
College grad	14	58.63	7.04

Table 3

Intercorrelations of Career Maturity Assessment (CMA) with My Vocational Situation (MVS), Gender, Age, and Educational Level

	MVS	Gender	Age	Ed Level
CMA	.78**	.10*	.14**	.17**
MVS		.08	.13**	.11*
Gender			.05	.18**
Age				.29**

*p < .05
**p < .01

Discussion and Recommendations

The CMA demonstrated several characteristics that support its potential usefulness as an indicator of career maturity. Reliability is .85. Correlation with the MVS indicates that it probably measures the same construct and that both concurrent and construct validity could be inferred--to the extent, at least, that they can be inferred for the MVS. Holland et al (1980) describe the MVS as having "a moderate degree of reliability and promising validity" (p. 1). The latter is described as the construct validity of the three MVS scales--Vocational Identity, Occupational Information, and Barriers. For total MVS score for the sample tested with both the MVS and the CMA, alpha was .88. The two instruments are intended to measure the same kinds of vocational needs or problems on the part of the respondent. Since the CMA score distribution was negatively skewed, with the mode higher than the median and the median higher than the mean, one might conclude that the items were not too difficult for most recruits to read and understand.

There are preferable ways of determining the face validity of the items and of setting the cutoff score. Budget and work-time restrictions, however, precluded such activities as the use of judges for determining face validity; administration of a longer, better-researched criterion measure than the MVS; factor analysis to study the construct validity of the CMA; and use of a standard-setting procedure such as the Angoff, Nedelsky, or Ebel methods (Livingston & Zieky, 1982) for establishing the cutoff score. These activities are suggested as recommendations for future, continued development.

In addition, the reliability of the dichotomous decision will need to be obtained empirically after the cutoff score has been in operation for an adequate period of time. The decisions made at various points above and below the cutoff will need to be examined to determine whether those who really need help with vocational decisions are being identified and whether false positives and false negatives are at an absolute minimum.

Furthermore, since the CMA has no subscales that might indicate specific problem areas (as in the MVS), there is a need to identify the kinds of problems that individual recruits have with regard to vocational

decision-making--for example, obtaining a better picture of their abilities and interests, locating occupational information, clarifying their goals, resolving value conflicts, or improving their decision-making skills. It may be possible to identify such problems through the CVG program. Recruits should also be given the option of asking for help if they feel they need it, regardless of whether they are above or below the CMA cutoff score. In fact, keeping track of such requests might provide still another way of checking on the accuracy of decisions based on the cutoff.

One last word should be said about setting the cutoff score at the median: Fortunately, the decision involved is not capable of having adverse impact on those below the median. The score will not be used for employment or promotion decisions. Its use should be entirely benign, both from the service's and the individual recruit's point of view; and, of course, it should be experimental until such time as policy and further research sanction its operational use.

Note: Financial support of this work by Battelle Scientific Services Program, in partial fulfillment of contract DAAG-81-D-0100, is hereby acknowledged. The views, opinions, and/or findings contained in this report are those of the author and should not be construed as an official Department of the Navy position, policy, or decision, unless so designated by other documentation.

References

Crites, J.O. (1973). Theory and research handbook: Career Maturity Inventory. Monterey, Calif.: McGraw-Hill.

Ginzberg, E. (1984). Career development. In D. Brown, L. Brooks & Associates, Career choice and development (pp. 169-191). San Francisco: Jossey-Bass.

Holland, J.L., Daiger, D.C., & Power, P.G. (1980). My vocational situation. Palo Alto, CA: Consulting Psychologists Press.

Livingston, S.A., & Zieky, M.J. (1982). Passing scores. Princeton, N.J.: Educational Testing Service.

National Assessment of Educational Progress (1976). The first national assessment of career and occupational development. Career and occupational development report no. 05-COD-00. Denver: Author.

Osipow, S.H. (1983). Theories of career development (3rd ed.). Englewood Cliffs, N.J.: Prentice-Hall.

Super, D.E. (1953). A theory of vocational development. American Psychologist, 8, 185-190.

Super, D.E. (1957). Psychology of careers. New York: Harper.

Super, D.E. (1974). Measuring vocational maturity for counseling and evaluation. Washington, D.C.: National Vocational Guidance Association.

681

NAVY R&D IN SUPPORT OF THE ARMY JOIN SYSTEM: LEVERAGING THE GOVERNMENT RESEARCH DOLLAR

Herbert George Baker,
Bernard A. Rafacz,
and
William A. Sands

Navy Personnel Research and Development Center
San Diego, CA 92152-6800

Inter-service competition and the consequences of rivalry among the various branches of the armed forces are legendary. Nevertheless, there are instances of highly productive, cooperative efforts that conserve precious research dollars and shorten the time line between conception and implementation. An excellent example is the research and development conducted by the Navy Personnel Research and Development Center (NAVPERSRANDCEN) in support of a new computerized system being developed by the Army for its recruiting forces.

Background

For nearly a decade, NAVPERSRANDCEN has held a leading position in the development of prototype automated systems that have direct applicability to military recruiting. Previous efforts of this research laboratory have resulted in a number of innovative products such as mobile van-based computerized military occupational information systems, complex algorithms that predict entry level job openings for periods as long as three months, and the Navy's automated classification and assignment system.

In the late 1970's NAVPERSRANDCEN began work on a sophisticated computerized system for the Navy Recruiting Command (NAVCRUITCOM). This was the Navy Personnel Accessioning System (NPAS) (Baker, Rafacz, & Sands, 1983a). It addressed a broad range of concerns at the recruiting station level; i.e., at the most forward terminus of recruiting operations. Subsumed within NPAS were components that harnessed the power of the microcomputer to the tasks of applicant screening, vocational guidance, assignment prediction, and recruiting management support (Baker, 1983; 1985b).

A demonstration version of this system was completed in 1981. Notwithstanding the successful development of this prototype, funds were unavailable for further prosecution of this effort. In addition, NAVCRUITCOM had no immediate plans for automating its recruiting station functions. Consequently, the NPAS project was terminated. Nevertheless, the body of expertise that had been built up in the course of developing the NPAS system held excellent potential for application in other arenas of personnel systems research.

Gearing up for the aggressive recruiting that will characterize the next several years, the Army launched a bold, innovative program to streamline and improve its recruiting and accessioning methods. The United States Army Recruiting Command (USAREC) is the proponent and developing agency for the Joint Optical Information Network (JOIN) System which is in the forefront of this effort (Bryan, 1982). Research and technical advisory services are being provided by the

U.S. Army Research Institute for the Behavioral and Social Sciences (ARI).

JOIN, a stand-alone microcomputer-based system, is in many respects similar to NPAS, while being technologically superior by virtue of advanced microcomputer capabilities. It has been implemented in Army recruiting offices nationwide. In the course of the recruiting process, Army enlisted applicants directly interact with this state-of-the-art accessioning system.

When the Army became the first to apply the benefits of automation to the front line recruiter, the opportunity for rapid inter-service technology transfer became apparent. Research managers of both laboratories became aware of a mutual R&D opportunity; one that could more swiftly advance the Army's project, while keeping intact a Navy research group with a proven track record. Accordingly, ARI proposed that the NPAS research team work on the JOIN System under ARI funding. The vehicle for this inter-service cooperative effort was an inter-laboratory agreement which provided for a 3-year effort beginning in FY82, and terminating on 30 September 1984 (Sands, Gade, & Bryan, 1982).

Task Areas and Products

Five major task areas provided the focus for the efforts of the Navy research group. Both in-house and contractor resources were enlisted in the ensuing work. Some of the products resulting from this project were directly adaptable from previous Navy innovations, while others represent new endeavors on the part of both laboratories.

Aptitude and adaptability screening was a task area where significant accomplishments were soon evident. The paramount result was the development of the capability to administer, score, and interpret a computer-based adaptive test for screening enlistment applicants. The specific product was the Computerized Adaptive Screening Test (CAST), which was designed to replace a conventionally administered paper-and-pencil instrument. Its purpose is to indicate probable success or failure on a comprehensive test battery used for applicant selection by all the armed services.

Test items had previously been developed under a NAVPERSRANDCEN contract with the University of Minnesota. Algorithms for item calibration, adaptive item selection and administration, scoring, and interpretation of test results were either developed or adapted from those extant in the psychological literature. An experimental version of the test had been completed for NPAS (Baker, 1983a; 1984a); and it proved readily portable to the Army's JOIN system.

Computer software to support the administration, scoring, and interpretation of the CAST on a developmental microcomputer system was written and documented (Baker, Rafacz, & Sands, 1983b). Subsequent to field testing and data collection (Sands & Rafacz, 1983), data analyses and instrument refinement were completed. CAST was refined and programmed to operate on the operational microcomputer system. When it was implemented nationwide on the JOIN

System, CAST became the first large scale operational use of computerized adaptive testing (Sands & Gade, 1983).

Thereafter, NAVPERSRANDCEN researchers continued to serve in a consulting role, assisting in program modifications required to interface the CAST with existing computer programs in the JOIN system, and modifications that permit on-line data gathering for further test validation.

Also developed under this task area was an instrument designed to assess military adaptability. This instrument is based on a values congruence approach in contrast to the more common method of empirically linking biodata with tenure. It was pilot-tested on a sample of recruits (N=540). The result was the identification of four items that discriminate between adaptables and nonadaptables. While further validation is indicated, an instrument of this type could be inserted into the automated screening procedures with little impact on applicant processing time or on the recruiter-applicant interaction.

A second major research focus was vocational guidance. The objective here was to conduct preliminary research in the design of a computerized vocational guidance (CVG) system that would be compatible with the Army recruiting environment. NAVPERSRANDCEN experience in developing interactive, user-friendly guidance systems proved to be a valuable and transferable resource.

Literature review, interviews with system developers, and hands-on appraisal were combined in a survey of available CVG systems, assessing their adaptability to Army recruiting purposes. The study produced preliminary indications of the components that would be needed in a recruiting-oriented CVG system. It concluded that it would be infeasible to adapt any extant system, and recommended development of an automated system specifically tailored to the needs of military recruiting and accessioning (Baker, 1984b).

Subsequently, work on a conceptual model of a recruiting compatible CVG system was initiated. A major step in developing such a system is detailing the organizational and operational constraints that would impact its design. Consequently, a study was undertaken and its results reflected in a report addressing these issues (Baker, 1985a).

Additional studies in vocational interests and values and their assessment delineated the many issues and controversies surrounding these areas of psychological concern, as well as the availability of appropriate instrumentation. Efforts to identify a suitable vocational interest inventory for administration to Army applicants at the recruiting station led to recommendations for the Vocational Interest Career Examination (VOICE) (Alley, 1978) as the most suitable inventory available.

An instrument to assess the career maturity of enlistment applicants was also developed (Diamond, 1985). In use, this instrument will determine whether an applicant requires more or less guidance in occupational exploration. It will also indicate to the recruiter the strength of applicant job preferences, thereby

assisting in focusing the sales interview with applicants whose strong enlistment motivation is job training. The instrument is being considered for incorporation into a JOIN-based system that is being designed to provide Army job information to students in community colleges.

Attention was directed within this task area to the preliminary classification of Army entry level occupations according to the Holland coding schema (Holland, 1973), and to a study of the feasibility of using expressed preferences in a recruiting-oriented CVG system. The occupational classification will facilitate job exploration and the elicitation of preliminary job choices. Expressed preferences may prove to be a method for rapidly focussing the classification interview, as well as demonstrating that applicant preferences are in fact being considered.

Work in a third task area, recruiting management support, sought to enlist the computer in lightening the recruiter's clerical task load. NAVPERSRANDCEN had developed several automated forms and reports generation capabilities as part of the NPAS effort (Baker et al., 1983a). Here again, technology transfer opportunities were obvious.

Interviews with Army recruiters and recruiting managers, and previous experience in the field were employed in assessing needs for forms generation, reports generation, and general word processing. Preparation of the Application for Enlistment (DD Form 1966), a multipage document, is a major administrative burden within the recruiting process. Information needed for creating an applicant file of data to produce the DD Form 1966 was determined and functional requirements for a computer printer which would generate the form were specified. Interactive computer dialogues, and complementary printout capabilities for the DD Form 1966 were developed, tested, and demonstrated.

To enable Army automation specialists to begin the task of porting the software to the operational system, documentation was expedited and delivered to ARI and USAREC concurrently. A program for an experimental "free-form" automated Application for Enlistment (capturing the same personal data as the DD Form 1966 but printing out on plain paper) was also developed and documented.

Sales technology is an integral, indeed vital, part of the recruiting process. The objective in this fourth task area was to develop an automated methodology for assessing enlistment motivation. Interviews with recruiting managers led to selection of the paradigm already employed in the Army recruiter's sales package as the basis for instrument development.

The product of this task area is a formalized technology for employing an established Army sales technique. When automated, and incorporated into a recruiting sales presentation on the JOIN system, this instrument would determine an applicant's dominant buying motive (DBM), i.e., the strongest motives toward enlistment. This information would allow the recruiter to focus on the appropriate features and career benefits, quickly tailoring the sales presentation to the applicant.

A fifth and final task area was that of personnel assignment. Using expertise gained in the development of NPAS and other automated assignment systems, coupled with information gained through interviews with Army recruiting and automation managers, the research team determined the required developmental efforts to interface the JOIN system with extant and planned Army-wide computer networks. This study highlighted the problems inherent in developing an assignment-prediction system, one that would provide information on entry level job openings and produce more optimal person-job matches.

Conclusions

This inter-laboratory cooperative effort proved very fruitful. Several applicant assessment instruments were developed, tested, and refined. One is already in operational use in the recruiting milieu, and in itself represents a milestone in computerized adaptive testing. Other instruments are being evaluated for further validation and possible implementation. Studies that are critical to the design of recruiting compatible automated systems were completed and their results made available to the research and operational communities through a number of working papers, technical reports, professional papers, and journal articles. A number of software packages were completed.

Necessary groundwork has been accomplished in a wide variety of research areas: on this base, the Army can develop enhanced capabilities for the JOIN system more rapidly. Because the recruiting operations of all the armed services are similar, products of this effort have high potential for further inter-service technology transfer. Endeavors such as this are exemplars of a developing trend within military personnel psychology R&D: maintenance of service-specific expertise bases, allied with decreasing research parochialism (Wiskoff, 1985). In sum, this project dramatically demonstrated the value of inter-service cooperation, an excellent example of the leveraging of the government research dollar.

References

Alley, W. E. (October 1978) Vocational Interest-Career Examination: Use and application in counseling and job placement. Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Baker, H. G. (May 1983) Navy Personnel Accessioning System (NPAS): II. Summary of research and development efforts and products. (NPRDC SR 83-35). San Diego: Navy Personnel Research and Development Center.

Baker, H. G. (January 1984) Computerized Adaptive Screening Test: Development for use in military recruiting stations (NPRDC TR 84-17). San Diego: Navy Personnel Research and Development Center.a

Baker, H. G. (February 1984) Computerized vocational guidance (CVG) systems: Evaluation for use in military recruiting (NPRDC TR 84-21). San Diego: Navy Personnel Research and Development Center.b

Baker, H. G. (July 1985) Designing a vocational guidance system for military recruiting: Problems and prospects - I. Organizational and operational considerations (MPL TN 85-7). San Diego: Manpower and Personnel Laboratory, Navy Personnel Research and Development Center.a

Baker, H. G. (1985) A prototype computerized vocational guidance system for Navy recruiting. Journal of Computer-Based Instruction, 12(3), pp. 76-79.b

Baker, H. G., Rafacz, B. A., & Sands, W. A. (May 1983) Navy Personnel Accessioning System (NPAS): III. Development of a microcomputer demonstration system (NPRDC SR 83-36). San Diego: Navy Personnel Research and Development Center.a

Baker, H. G., Rafacz, B. A., & Sands, W. A. (May 1983) Initial development of a computerized adaptive screening test (CAST) for use in military recruiting. Paper presented at the annual conference of the International Personnel Management Association (IPMAAC), Washington, DC, 22-26 May 1983.b

Bryan, J. D. (October 1982) Recruiting evolves toward tommorrow. All Volunteer. Fort Sheridan, IL: U.S. Army Recruiting Command.

Diamond, E. E. (July 1985) Development of the career maturity assessment (MPL TN 85-7). San Diego: Manpower and Personnel Laboratory, Navy Personnel Research and Development Center.

Holland, J. L. (1973) Making vocational choices: A theory of careers. Englewood Cliffs, NJ: Prentice-Hall.

Sands, W. A. & Gade, P. A. (1983) An application of computerized adaptive testing in U.S. Army recruiting. Journal of Computer-based Instruction. 10, (3&4), 87-89.

Sands, W. A., Gade, P. A., & Bryan, J. D. (1982) Research and development for the JOIN system. Paper presented at the 24th annual conference of the Military Testing Association, San Antonio, TX, 1-5 November.

Sands, W. A. & Rafacz, B. A. (1983) Computerized adaptive testing (CAT) for the U.S. Army JOIN system. Paper presented at the 25th annual conference of the Military Testing Association, Gulf Shores, AL, 24-28 October.

Wiskoff, M. F. (August 1985) Military psychology and national defense: Making an difference. Division 19 Presidential Address delivered at the 93rd annual convention of the American Psychological Association, Los Angeles, CA, 23-27 August.

Vocational Interests as Predictors of Army Performance[1]

Hilda Wing
U.S. Army Research Institute

Bruce N. Barge and Leaetta M. Hough
Personnel Decisions Research Institute

In: Elements of a Military Occupational Exploration System
Military Testing Association, October 1985
San Diego, California

Measures of vocational and occupational interest have been used in selection for Army enlisted occupations for many years. In this paper we will describe how such measures have been used in the recent past, review current Army research which will link such measures to performance in Army jobs, and identify critical issues that must be resolved in order for interest measures to be effective in a selection and classification program.

The Army's Use of Interest Measures in Selection/Classification

Use of vocational interest measures for classification into Army training was part of the Army's selection and classification for enlisted personnel from 1958 until 1980. The Army Classification Batteries, followed by the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7, included forms of the Army Classification Inventory (ACI), which contained sentences describing activities with which an applicant could agree or disagree. Four scale-scores were obtained from each applicant: Combat, Administrative, Mechanical, Electronics. These scale-scores were incorporated with ASVAB cognitive ability subtest scores to produce Aptitude Area (AA) composites. For example, the Combat AA included both ability and interest measures. Empirical data supporting this use had been provided by the developers of ACB-73 (Maier & Fuchs, 1972). Interest measures were dropped from the Army enlisted classification system with the introduction of new ASVAB forms in October of 1980.

The Army's current Project A is, among other things, the largest selection and classification research effort to date. The initial function of Project A is to validate the ASVAB against Army performance. An additional aspect of Project A's mission is to develop new predictors which will cover attributes that the ASVAB does not. ASVAB is more than adequate for selection into Army training (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984). What we are more concerned about is classification and, in addition, performance on the job, successful completion of the first tour, and reenlistment eligibility. To that end there has been

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Deparment of the Army.

developed an evolutionary model of predictor space. This model conceives of predictor space as having three components. First, the cognitive-perceptual component includes measures of verbal, quantitative, and spatial abilities. Next, the perceptual-psychomotor component includes perceptual speed and accuracy, short-term memory, multi-limb coordination, and movement judgment. We have developed a mini-battery for this component which is administered on an IBM-compatible microcomputer with a custom-designed response pedestal. Finally, the non-cognitive component covers both biographical/temperament (personality) and vocational interest measures.

The evolution of this model of predictor space has been firmly anchored to data, as follows. Project A has so far completed research on a first cohort of Army enlisted personnel, those entering in FY 1981 and 1982. The second cohort includes those soldiers who entered the enlisted service during FY 1983-1984. It includes both longitudinal and concurrent components; the longitudinal is included in the concurrent. For the longitudinal effort, we developed our first test battery, the Preliminary Battery, from readily available, off-the-shelf paper and pencil measures of cognitive and non-cognitive attributes. We administered the Preliminary Battery prior to training, to soldiers in four selected MOS, from October 1983 through June 1984. This year we obtained measures of training success and early attrition for this sample.

This summer, we are testing the concurrent component of this 1983-84 cohort with a second, new battery, in conjunction with a full complement of performance measures. We have added another 15 MOS, and the perceptual-psychomotor component of predictor space is being evaluated with micro-computers. Data collection should be complete by late November, 1985. While we have no analyses completed for what we are calling the Trial Battery, we do have some information about its immediate fore-runner, the Pilot Trial Battery.

A complete longitudinal effort is planned for the FY 1986-1987 cohort, to begin sometime next year. There will be the Experimental Battery, which will be much like the Trial Battery, to be administered to soldiers entering training in each of our selected MOS. Subsequently, we will be administering the appropriate performance measures to these soldiers. At the same time, we also plan to evaluate the performance of second-tour members of our 1983-1984 cohort.

Results for the Preliminary Battery

The Preliminary Battery included the Air Force Vocational Interest Career Examination (VOICE), which assesses 18 basic interests (Alley & Matthews, 1982). Because of the research on the Holland hexagonal model of vocational interests, we investigated its appropriateness. We factor-analyzed both the items and scales of the VOICE. We were able to recover the 18 basic interest scales quite nicely from the item factor analyses (Hough, Dunnette, Wing, Houston, & Peterson, 1984). We were able to find the Realistic group of occupational interests, but all the others clumped mostly into one group. In hindsight this made perfect sense. The majority of occupations in the enlisted military service are Realistic in

nature, as they are jobs in the skilled trades. There are a handful of Investigative occupations, some Conventional, and some Social occupations. For virtually no occupation in the enlisted ranks does the Artistic or Enterprising label fit.

What evidence was there of criterion-related validity for these interest measures? Available criteria were existing training grades and early attrition (status as of December 1984, or an average of one year of service). For training, the cognitive tests of the Preliminary Battery appeared to have some predictive power, although the coefficients were not large and not much larger than those obtained for the ASVAB. The attrition analyses are currently incomplete. This criterion will be especially hard to predict because the early attrition was fairly low, about eight percent. While some of the VOICE scales were significantly related to attrition in each of the four MOS, the correlations were quite low. The coefficients for some of the biodata/temperament scales, which evaluated aspects of socialization, were higher than those for interests. The domain of causes for discharge in the Army extends from "disciplinary" through "for good of service" to "unsuitable unknown." It is likely that early attrition in the Army, particularly that through the Trainee Discharge program, may be more disciplinary than anything else. Thus, the predictiveness of the socialization scales is understandable.

The VOICE scales were not related to any great extent with the other measures evaluated, including the ASVAB. It is likely that as various criteria mature (later attrition, re-enlistment) or are administered as part of the Project A data collection (commitment, effectiveness), these early measures of vocational and occupational interests will have a better chance to demonstrate what they can do.

Results from the Pilot Trial Battery

The Pilot Trial Battery was field tested during the fall of 1984. Soldiers supplied data to evaluate the properties of the battery, including test-retest stability. We called our interest measure here the "Army VOICE," or AVOICE. We obtained this by starting with the VOICE, cutting back items on most of the 18 scales while adding scales for Army interests which are not duplicated in the Air Force, such as Infantry, Armor/Cannon, Science/Chemical Operations.

Psychometrically, the new instrument worked well, except that the factor analyses yielded the same pair of factors as before. For the Pilot Trial Battery, these factors appeared to be described better as "Combat" and "Combat Support," rather than "Realistic" and "Non-Realistic." This is a matter of taste rather than substance, as there is confounding of terms. The Combat occupations are Realistic while the Combat Support occupations cover the other five corners of Holland's hexagon. But, this is, we judge, the occupational reality of the Army enlisted world. The reliability and stability of the interest scales were excellent, in the .80's and .90's. There were no performance criteria available for this sample, but we did inspect the overlap of the interest measures with the remaining components of the Pilot Trial Battery and the ASVAB. The intercorrelations between AVOICE scales and other scales were generally low.

Preparing the Trial Battery from the Pilot Trial Battery consisted mainly of cutting back, so that a 6-7 hour battery was reduced to one requiring less than four hours. The AVOICE in the Trial Battery being administered now includes 176 items and takes about 15-20 minutes to administer. It will provide scores for interests in 22 Army occupations.

Issues in the Operational Use of Interest Measures in Selection and Classification

We see at least five major issues to be confronted in determining when and how to use measures of vocational interests in selecting and classifying for military enlistment. The first four are clearly technical while the last is more of a policy issue which can be informed by our technology.

First, the complete hexagonal model of Holland's vocational interest theory appears to be inappropriate for predicting performance in Army occupations. We tend to forget the context-sensitivity of models. The domain of Army jobs maps onto only a portion on the theorized hexagonal interest space, mainly that corner called Realistic. All of the other Army jobs, which could be characterized as involving interests from the Investigative, Social, and Conventional corners, appear to clump together. At this time we do not know whether this simple differentiation will provide all the predictability possible, given the available criteria, or whether further distinction into occupational scales will be warranted. But, it is clear that approaches using a complete Holland model will have limited applicability for the spectrum of Army enlisted occupations.

Second, the selection of appropriate criteria for vocational interests to predict is a major concern. Should criteria be those we consider as maximal effort, such as job knowledge tests and hands-on measures? Or should they be typical effort types of measures, such as motivation? We really need to know more about these criteria. One of the goals of Project A is to improve our conceptual understanding of the criterion space. This is clearly a worthy and necessary goal.

Third, how should predictors and criteria be used? The primary function of any interest measure is to direct the individual towards some occupations and away from others. That is, the object is classification. Regardless of the specific criteria, there are questions about the form of the predictors to use. Should we use scores from occupational scales, or should we use factor scores? Should we use single scores, or do we need to investigate configurations, or profiles? How should we combine interest measures with measures from other domains, such as the cognitive? It could be that positive interest in a specific area can compensate, to some extent, for lower ability for that area (Matthews, 1982). What are the characteristics of the sample sizes, the psychometric properties of the measures, that must be present for us to be able to make any kind of definitive statement concerning such claims?

Fourth, what exactly are we trying to predict: Success or avoidance of failure? This is the more complex issue concerning the fact that the

Army, and perhaps most employers in general, cannot always use people in what those people are best at. For example, one of the MOS in Project A is the Combat Medic. We have administered a complete battery of performance measures to several hundred Combat Medics so far in addition to the Trial Battery. However, at this point in time the United States is currently not in any general armed conflict, and there is little opportunity for these soldiers to practice their training in any realistic environment. Some of them may be working in maternity wards while others spend most of their time in the motor pool. We find it difficult to understand exactly how an interest in medical activities, absent other information, will be predictive of important criteria for these soldiers. Other examples are possible. How should interest measures be used in such cases?

The fifth and final issue concerns where in the enlistment process is it most appropriate to use interest measures? In the All-Volunteer Army, they may be more appropriately used by the recruiter. Should they be used in a mandatory or advisory way? Perhaps this is a technical question as much as are the other four: Are interest measures more predictive, of whatever criteria we can come up with, in whatever psychometric fashion determined effective, when these measures are used in an advisory fashion rather than a mandatory one?

This report has provided a brief description of how the Army is investigating the use of vocational interests in predicting performance in Army jobs. Project A will be providing vast amounts of data which will better inform our use of these measures. However, this use may be complex. The empirical data will, we trust, point us towards better use.

References

Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination: A description of the instrument and possible implications. Journal of Psychology, 112, 169-193.

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J., & Peterson, N. G. (1984). Covariance analyses of cognitive and noncognitive measures in Army recruits. Paper presented at the convention of the American Psychological Association, Toronto, Ontario, Canada.

Maier, M. H., & Fuchs, E. F. (1972). Development and evaluation of a new ACB and aptitude area system (Technical Research Note 239). Alexandria, VA: U.S. Army Research Institute.

Matthews, M. D. (1982). Vocational interests, job satisfaction, and turnover among Air Force enlistees. Paper presented at the Fourth Annual Learning Technology Congress and Exposition, Society for Applied Learning Technology, Orlando, FL.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative Armed Services Vocational Aptitude Battery (ASVAB) area composites (Technical Report 651). Alexandria, VA: U. S. Army Research Institute.

Army Career Vocational Guidance As A Recruiting Tool

Allyn Hertzbach, Deirdre J. Knapp, and Richard M. Johnson
U.S. Army Research Institute for the
Behavioral and Social Sciences
Alexandria, Virginia 22333-5600[1]

What is Army recruiting doing in the vocational guidance business? Surely, the Army can find more efficient ways to meet its manpower needs. This point of view is not unreasonable and might find much support, especially among the other services. But the Army is considering the feasibility of providing career vocational guidance (CVG) via the JOIN system. This paper addresses the overall approach currently being pursued by the Army Research Institute (ARI) and the Army Recruiting Command (USAREC). First, the rationale for the project is discussed, and a description of the components of the vocational guidance package follows. The feasibility research now being conducted with the American Association of Community and Junior Colleges (AACJC) is briefly described and is followed by a look at what future research might be pursued.

Program Rationale

Anyone familiar with career vocational guidance knows that automated career information delivery systems (CIDS) are available in many schools and employment centers. There are states that have such systems available at little or no cost to residents. In fact, the military services are considering providing software that would allow high schools to use the Armed Services Vocational Aptitude Battery (ASVAB) as the ability measure for the most often used automated delivery systems. There is even some consideration being given to developing a DoD CIDS that could be deployed across the country to encourage young people to take the ASVAB and to consider joining one of the military services.

The Army began planning to develop its own service specific vocational information system before DoD began to examine the possibility of providing its own program in career vocational guidance. And the Army is still considering the feasibility of an automated system of its own. The main reason for continuing this work is simply that the Army has the most difficulty of all of the services in meeting its manpower needs. Another compelling reality is that the recruiting environment is becoming increasingly competitive in that the Army has to vie with the other services, as well as an increasing number of private sector employers. A third reality is that the pool of the nation's young people is shrinking and will not begin to increase until the mid 1990s. Finally, the Army has the automated means (JOIN) to prepare a tailored sales package that would greatly benefit its recruiters, not to mention

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army

other possible deployments of the system, as in high schools or two year colleges.

Aside from the sales advantages to be derived from improving the JOIN sales package with career information and vocational measures, the Army recruiter would be able to direct or funnel applicants more systematically by using occupational groupings rather than being asked about specific jobs and and having to soft pedal or evade the issue of specific jobs or particular kinds of job training. The Vocational Guidance package would have available to it groupings of skill clusters developed by the Army Recruiting Command, rather than particular military occupational specialties, so that the prospect could view more generalized job information. Therefore, the applicant could see a realistic preview of relevant jobs via JOIN videodisc presentation without also being predisposed to a particular job, upon which he might predicate enlistment. If the other services and DoD approved, the Army might also be interested in including the Military Occupational And Training Data (MOTD) as another resource for its vocational guidance package, which collapses military jobs across the services and does a crosswalk with civilian jobs.

Anatomy of the Army Career Vocational Guidance (CVG) Package

The CVG package being developed for the JOIN system will be compatible with the current JOIN software, replacing some components and adding other components. A brief description of the JOIN system is included for those readers unfamiliar with the system. The Joint Optical Information Network (JOIN) is a Z80 based microprocessor with 64k RAM, I/OS operating system with two double-sided five and one quarter inch floppy disc drives. The JOIN system includes a keyboard, color monitor, dot matrix printer, modem, and videodisc player. The videodisc player is under random access control through the JOIN system. The system is employed by a recruiter with a prospect to determine the individual's needs, interests, wants, as well as to determine the prospect's likelihood of qualifying for enlistment. The system also presents information about Army jobs, benefits, and enlistment options. The JOIN system is currently deployed at almost all Army recruiting stations.

The Career Maturity Assessment (Diamond, 1984) is being considered for inclusion in the Army's CVG package. The instrument was developed for the Naval Research and Development Center and provides a score that indicates whether an individual is ready to make reasonable career decisions or not. Though there is disagreement about what the actual career maturity dimensions are, Super (1974) suggests that career maturity is the readiness to make reasonable career decisions at given decision points in career development. He describes career maturity as the congruence between a person's vocational behavior and the behavior expected at that person's age. Diamond (1984) reports that self-knowledge, self-concept, and decision making skills appear to be the common elements in research aimed at describing and explaining career maturity. Career maturity information can be very useful for both the prospect and the recruiter. This information should be valuable to the recruiter and help him to recognize less stable decision making potential. The

recruiter might want to take more time to clarify needs and interests with individuals who do not meet the cutoff score on the instrument and expect more fluctuation or poorer selections from "immature" individuals. "Mature" individuals, of course, are not immune from mistakes or fluctuations, especially at the ages when Army recruiters are likely to see them.

Two interest inventories are being considered for inclusion in the CVG package. Both of these instruments are currently being developed by ARI. One is an in-house effort, the Job Interest Survey(JIS), and the other, the Vocational Interest Profile (VIP), is an ARI contract effort (Faust, Unger, and Schmitz, 1985). Test development for these instruments continues and is not yet far enough advanced to speculate about which will be selected. One possibility being considered is combining the best items of the two tests and automating the result. Thus far, only preliminary testing of the instruments has occurred (Faust, Hertzbach, and Knapp, 1985) with a small sample of 160 soldiers. This pilot effort was implemented to work out the administration, to test the reading level of the instruments, to determine how long it took soldiers to take the test, and to determine the clarity of the instruments. The two instruments were administered to new recruits at Army reception stations as part of the 1985 New Recruit Survey. Thus, more extensive analyses will be performed to determine the potential usefulness of the instruments.

The next element of the CVG is job information, and there are a number of possibilities. The JOIN system is particularly well equipped for providing large reference resources. With a few 5 1/4" diskettes very large amounts of information can be made available to the prospect and recruiter. There are several sources of information which could be useful, but the most important consideration, short of accuracy and comprehensiveness of the job information, is the policy that the recruiter cannot, indeed, must not, sell specific jobs. With this restriction in mind, the skill cluster groupings developed by USAREC and Career Management Fields (CMF, groups of MOSs) are more appropriate than is a directory of all of the specific Army jobs (MOS). Another excellent source of job information is the Military Occupational and Training Data software developed for joint service use. The difficulty encountered in using this software is the number of diskettes that would be required to add it to the package. Increasing the number of diskettes makes the package a bit more awkward to use, not to mention increasing the chances for losing or impairing the materials. Despite these considerations, the MOTD software has many advantages and with the approval of the joint service oversight committee could add a great deal of useful information to the Army's CVG package. The available software for including realistic and comprehensive job information is an important part of the CVG and should greatly enhance its sales utility.

Aside from job information, there are a number of incentives and Army programs about which prospects should be informed. JOIN is an excellent medium for presenting this kind of information. There are many varieties of information that can be tailored for specific audiences. If the prospect is a college student or has some college credit, information about advanced grade for college credit or the Army College Fund or opportunities to become an officer (OCS and ROTC) should be presented. This part of the package furthers the sales dialogue and begins to focus on the reasons for joining and the issue of joining. With the acquisition of information and prequali-

fication (via the Computerized Adaptive Screening Test (CAST), the JOIN ability measure that predicts Armed Services Vocational Aptitude Battery (ASVAB) performance), a well-informed Army recruiter and the prospect can begin to discuss the tradeoffs for enlistment. If the prospect is likely to qualify mentally and physically, the general information can be personalized and presented in the most attractive way possible. The recruiter knows what job interests and needs the individual has and can link these to available incentives and programs. This kind of interaction could also largely be automated and allow the prospect to go through the package alone at his school counseling office.

CVG Feasibility

Now that the rationale and the components of the Army's CVG program have been presented, the practical consideration of whether or not this package can be deployed beyond the recruiting station becomes important. The Army needs to get to the high quality market and not just wait for that market to walk into the recruiting station. One market that the Army has never systematically approached is the two year college market, and these schools seem to be an excellent place to develop and test the feasibility of the Army's CVG. ARI currently has a contract with the American Association of Community and Junior Colleges(AACJC) to determine if the CVG can function to the mutual advantage of the Army and the two year colleges. If the CVG is seen as appropriate, the package could also be employed, with some modification, in high schools, as well. However, the primary purpose for the CVG is for the Army recruiter in uncovering dominant buying motives and to structure the prospect's job expectations (without selling a particular job) and to match the prospect's needs and interests with Army opportunities.

Briefly, the goal of the contract effort with AACJC is to develop six sites at junior colleges to test and develop the CVG. Attempting to get an estimate of the potential market, targeting advertising to the market, discovering dominant buying motives, and identifying a tailored sales approach are particular aims of this work. JOIN systems are to be provided to each site, including the appropriate software and other materials. The best methods for implementing the CVG are also to be determined--stand alone or with the college counselor. Of interest also in this research effort is identifying barriers to enlistment and most effective ways of stating tradeoffs for the prospects and the recruiters.

This research effort is more an exploratory and qualitative project than a tightly designed experiment. Aside from the aforementioned purposes of the research, there is a need to establish good rapport with the two year schools, as well as with the AACJC. The schools and the AACJC need to understand that we are not trying to compete with them for their students, rather that we are providing ways for students to finish school and enabling them to further their education (Army College Fund, GI Bill). Having the schools collect and provide critical commentary on the CVG components, testing the CVG's reception at the schools, and learning how best to interface the Army recruiter with the schools are qualitative issues that are also being addressed. This project will be completed at the end of fiscal year 1986.

Discussion

The Army Career Vocational Guidance research project will assess vocational interest and, perhaps, career maturity; assess ability (CAST); provide realistic military occupational information; provide incentive and program information; and articluate clear and attractive tradeoffs to encourage prospects' decisions to enlist. The strategy for implementing this program is to make it available to Army recruiters on JOIN and to provide it to institutions (e.g., two year colleges and high schools) for the benefit of their students. The students will receive vocational information about themselves, as well as information about the Army opportunity. They will also learn about military occupations and the training required for these occupations. This service will be provided at no cost to the schools or the students. The Army recruiter will receive feedback in some form from the schools using these packages. Perhaps, schools will send prospect lists of students who express some interest in learning more about the Army.

The actual viability of an Army CVG is, at this point, undetermined. There is no question, however, that the Army needs to explore all avenues that might produce high quality prospects in this difficult recruiting environment. However, this effort can only be made to good effect with the cooperation of the other services. If there is an all service effort, the Army CVG program will, of course, need to be made compatible with whatever is implemented, including the possibility of scrapping any school deployment. However, the information that is learned from our current effort with the AACJC should be useful in both the Army's effort, as well as the all service, DoD effort.

As previously suggested, the future of the Army Vocational Guidance package depends in large part on it viability in the schools and to the extent that it can be coordinated with all service efforts. But, assuming that the program is viable, the next phase of the development would be to design a comprehensive empirical research effort to test the hypotheses developed from the current effort with the AACJC and build the package, implement it, assess its effectiveness, fix it, and deploy it as policy makers deem appropriate.

References

Diamond, Esther E. (1984). Development of the Career Maturity Assessment (Draft Report). San Diego, CA: Naval Personnel Research and Development Center.

Faust, D., Unger, K. & Schmitz, E. (1985). Development of a Prototype Army Vocational Interest Inventory for the Enlisted Personnel Allocation System (Tech Report). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Faust, D., Hertzbach, A. & Knapp, D. (1985). Field Trials of Two Vocational Interest Assessment Techniques: The Vocational Interest Profile (VIP) and Job Interests Survey(JIS) (Unpublished Report). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Super, D. E. (1974). Measuring Vocational Maturity for Counseling and Evaluation. Washington, D.C.: National Vocational Guidance Association.

Weltin M. (1984). Job Interests Survey (unpublished technical papers). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

C78

Developing New Attribute Requirements Scales for Military Jobs

Elizabeth P. Smith
U.S. Army Research Institute[1]
5001 Eisenhower Avenue
Alexandria, Virginia 22333-5600

Conducting empirical validity investigations to predict job performance is not always feasible. Even when empirical approaches are undertaken, such as the ongoing ARI Project A to improve the selection, classification and utilization of enlisted personnel, it is rarely possible to include all jobs within an organization. Given the complexities of empirical validation, it is necessary to develop other methods for matching people to jobs and optimizing their performance.

One approach is to obtain rational estimates of the human attributes (i.e., abilities, characteristics, and interests) which are required for successful job performance. When gathered systematically from qualified judges, these estimates can be summarized as profiles of required attributes. Then, measures of individuals' attributes can be matched to such profiles for selection and classification purposes. In addition, knowledge of required attributes is potentially useful for (a) designing new systems and training programs that are within the capacities of available personnel and (b) generalizing empirical validity data to new and different jobs, by grouping them on the basis of similarity of attribute profiles (Fleishman, 1982; Pearlman, 1980). The latter application is especially pertinent to the Army's Project A, which is collecting validity data for only 19 Military Occupational Specialties (MOS).

A well-researched method of determining ability requirements is the rating scale approach developed by Fleishman and his associates (see Fleishman & Quaintance, 1984 for a comprehensive summary), based on a taxonomy of 40 cognitive, perceptual, physical and psychomotor abilities. With these scales, a rater decides if an ability is necessary for errorless job performance, and, if so, estimates the level required on a 7-point, behaviorally-anchored scale.

Early outcomes from Project A provided an opportunity to develop a new set of rating scales based on a new taxonomy of human attributes. An expert judgment task (Wing, Peterson, & Hoffman, 1984) obtained estimates of validity for 53 predictors against 72 criterion constructs from 35 personnel psychologists. Factor analysis of the data yielded 21 clusters of the 53 cognitive, perceptual, psychomotor, temperament and interest predictor variables. A predictor test battery based on these 21 clusters has been developed and is being validated. The purpose of this paper is to discuss the initial construction and testing of a new set of scales for estimating job requirements which is based on these 23 clusters (hereafter called "attributes"). As more data become available, it is expected that the taxonomy of predictors (and test battery) may change. The rating scales will be revised to reflect these changes.

A set of scales based on the Project A taxonomy has several potential advantages over the Fleishman ones. The most salient feature is that obtained profiles of attribute requirements will directly correspond to

[1]The views expressed in this paper are those of the author and do not necessarily reflect the view of the U. S. Army Research Institute or the Department of the Army.

Project A validity data. It will include temperament and interest measures that are not among the Fleishman scales and will not include those attributes/abilities for which no predictor tests are given. Additional benefits (e.g., lower cost, more efficiency) may be possible with this set of scales. It was designed to be used by work supervisors rather than personnel psychologists and contains primarily Army-specific behavioral anchors with only about half as many attributes to rate as Fleishman's.

For any rating scales to be useful in practice, they must give reliable and valid scores. The effort reported here examined issues related to the reliability of the ratings. Validity investigations will occur later. The following issues were examined here. First, how closely do raters agree, i.e., how high is interrater reliability? Second, how well do the scales differentiate across attributes (i.e., yield non-flat profiles) within a job and across the attribute profiles of different jobs? Finally, can the scales be used to identify attributes for which differences in level of the attribute most influence performance? For some attributes, higher levels may be required for better performance whereas for others, once a minimal requirement is met, having a greater amount of the attribute has no additional effect on performance.

METHOD

Subjects. Thirty-six Non-commissioned Officers (NCOs) from the Cannon Crewman MOS and 39 NCOs from the Motor Transport Operator MOS, all males located overseas, participated as Subject Matter Experts (SMEs).

Instrument. The Attribute Assessment Scale, which was empirically developed for this research, consists of a set of behaviorally-anchored scales for 20 of the 21 attributes in the Project A taxonomy plus two additional attributes, Stamina and Physical Strength, which were thought to enhance face validity. A scale for Enterprising Interests was eliminated because it was impossible to generate items for this attribute which were sufficiently different from those falling under Self-Esteem/Leadership. The names of the attributes were modified from the original Wing, et. al. (1984) labeling for better comprehension by SMEs. The final instrument had one page per attribute. Below the definition at the top, there were three 7-point vertical scales, placed side-by-side, to enable three responses. A zero-point was added to indicate the attribute was not required at all. SMEs circled the number corresponding to the appropriate level for their job.

To construct the scales, comprehensive definitions for the attributes were developed so as to be readily understandable by people who were not trained in personnel research. A pool of items for potential anchors (i.e., behavioral statements) was generated. Ten items per attribute were ultimately selected, after screening by two to four other researchers. These were presented with the appropriate definition in an anchor-rating instrument. Initially, 26 NCOs from either the Administrative Specialist or Military Police MOS rated each item on the amount of the attribute represented by or needed for the behavior described. Items with mean ratings that were the highest, lowest, and closest to 4.0 (midpoint) that also had a standard deviation less than 1.5 were selected as scale anchors. Using these criteria, scales could be created for only 11 attributes.

After identifying difficulties related to (a) task comprehension, (b) response format, (c) failure of raters to differentiate effectively among items, and (d) a few of the definitions and items themselves, I revised the anchor-rating instrument and administration procedures, adding a

15-minute training period. This instrument was given to another sample of NCOs (N=28) from the same two MOS. From the second administration, using the criteria indicated above, three anchors were obtained for all but two of the attributes (Social Interaction and Stress Reaction for which only two anchors were selected) to form the Attribute Assessment Scale.

Procedure. SMEs rated the level of each of the 22 attributes that is required to perform Skill Level 1 (entry level) work under combat-readiness conditions in his own MOS for three performance levels: at the 15th, 50th, and 85th percentiles. In addition to the written instructions, SMEs received extensive training in how to complete the task, including a step-by-step demonstration of the actual rating process using the anchors as guides. Training and responses to questions took about an hour. Early ratings were checked to ensure comprehension of the directions before raters proceeded with the rest of the task. Ratings took about 30-45 minutes.

Analyses. Intraclass correlation coefficients (ICCs) were calculated from Raters X Attributes ANOVAs over all attributes and separately for the three major domains (i.e., cognitive/perceptual, physical/psychomotor, and noncognitive) for each of the three performance levels. The ICCs estimate the reliability of the mean ratings [r(k); k=number of raters], an index of interrater reliability. Also, an MOS X Attributes X Performance Levels univariate repeated-measures ANOVA was performed.

RESULTS

Eight Motor Transport Operators were eliminated from the analyses due to the logical inconsistency of their data. R(k) coefficients over all attributes were, in increasing order by performance level, .75, .77, and .69 for Cannon Crewmen (k=36) and .74, .74, and .69 for Motor Transport Operators (k=31). For the domains, r(k) coefficients ranged from .61 to .79 across performance levels and MOS. There were two exceptions to this: Physical/psychomotor reliabilities were very low for both MOS at the 85th percentile [r(36)=.13; r(31)=.38] performance level.

None of the effects involving MOS for the MOS X Attributes X Performance Levels ANOVA were significant. There were significant main effects for Attributes [F (21,1365) = 6.98; p = .0000] and Performance Levels [F (2,130) = 398.36; p = .0000] and a significant effect for the Attributes X Performance Levels interaction [F (42,2730) = 2.51; p = .0000]. Scheffe' comparisons between means within performance levels by MOS indicated significant differences between only the highest and lowest means, which ranged from 1.09 to 1.75. Means and standard deviations for all ratings are provided in Table 1.

DISCUSSION

In comparison to the very high Intraclass Correlation Coefficients (ICCs) obtained by Fleishman and associates or those discussed by Rossmeissl (1985) within this symposium, the ICCs from this research are weak, especially since around 30 raters are needed to obtain coefficients of at least .60. ICCs are based on variance components. As such, low (or uninterpretable) reliabilities result if there is too great a between-subjects variance and/or too little within-subjects variance. The low reliabilities obtained here appear to be a function of both. Previous research on ability assessment has found mean ratings that varied from very low (even "Not required") to very high (7) across attributes. This was not the case here. The inclusion of three performance levels may have

Table 1

Mean and Standard Deviations of Ratings of Attribute Requirements for Cannon Crewman and Motor Vehicle Operator MOS at Three Performance Levels.

Attributes	MOS[a]	Performance Level					
		15th Percentile		50th Percentile		85th Percentile	
		M	SD	M	SD	M	SD
Cognitive							
Verbal Ability	C	2.86	(.93)	4.17	(.91)	5.33	(1.10)
	D	2.87	(.83)	4.32	(.85)	5.32	(1.08)
Memory	C	2.44	(1.18)	3.89	(.89)	5.53	(1.11)
	D	3.26	(1.26)	4.26	(1.03)	5.16	(1.27)
Reasoning Ability	C	2.78	(1.31)	3.94	(1.17)	4.89	(1.33)
	D	2.45	(.96)	3.77	(1.02)	5.00	(1.32)
Number Facility	C	2.06	(1.12)	3.47	(1.16)	5.08	(1.52)
	D	2.48	(1.18)	3.90	(1.30)	4.94	(1.48)
Mechanical Comprehension	C	2.86	(1.36)	4.39	(1.18)	5.50	(1.08)
	D	2.97	(1.33)	4.23	(1.14)	5.20	(1.00)
Information Processing	C	2.50	(1.08)	3.58	(1.30)	4.81	(1.37)
	D	2.68	(1.30)	3.90	(1.08)	5.03	(1.17)
Closure	C	2.78	(1.33)	4.08	(1.25)	4.89	(1.41)
	D	2.68	(1.42)	3.87	(1.45)	4.61	(1.67)
Visualization	C	2.33	(1.1[illegible])	3.58	(1.30)	4.69	(1.51)
	D	2.29	(1.30)	3.42	(1.43)	4.26	(1.84)
Perceptual Speed & Accuracy	C	2.75	(1.52)	3.97	(1.38)	4.94	(1.31)
	D	2.97	(1.33)	4.16	(1.29)	4.71	(1.40)
Physical/Psychomotor							
Physical Strength	C	3.75	(1.32)	4.89	(1.14)	5.67	(1.17)
	D	3.58	(1.39)	4.61	(1.20)	5.32	(1.25)
Stamina	C	3.06	(1.45)	4.53	(1.11)	5.69	(1.19)
	D	2.68	(1.30)	3.94	(1.41)	4.84	(1.63)
Multilimb Coordination	C	2.80	(1.47)	4.20	(1.21)	5.26	(1.40)
	D	3.34	(1.54)	4.45	(1.24)	5.48	(1.12)
Dexterity	C	3.00	(1.35)	4.47	(1.08)	5.50	(1.00)
Non Cognitive							
Steadiness/Precision	C	2.83	(1.40)	4.08	(1.25)	5.47	(1.36)
	D	3.06	(1.29)	4.52	(1.09)	5.29	(1.07)
Social Interaction	C	3.14	(1.62)	4.44	(1.59)	5.34	(1.70)
	D	2.58	(1.71)	3.74	(1.44)	4.65	(1.70)
Stress Tolerance	C	3.03	(1.50)	4.22	(1.27)	5.12	(1.43)
	D	3.10	(1.47)	4.27	(1.34)	5.27	(1.20)
Conscientiousness	C	2.66	(1.24)	4.09	(.89)	5.31	(.99)
	D	3.19	(1.47)	4.35	(1.11)	4.97	(1.25)
Work Orientation	C	2.91	(1.46)	4.29	(1.18)	5.54	(1.17)
	D	2.90	(1.45)	4.16	(1.10)	5.39	(1.17)
Self Esteem/Leadership	C	3.00	(1.26)	4.25	(1.20)	5.47	(1.21)
	D	2.48	(1.55)	3.84	(1.37)	5.00	(1.41)
Athletic Ability/Energy	C	2.89	(1.35)	3.92	(1.16)	4.94	(1.19)
	D	2.87	(1.55)	3.73	(1.48)	4.33	(1.71)
Realistic Interests	C	2.54	(1.40)	3.71	(1.25)	4.94	(1.66)
	D	2.48	(1.12)	3.61	(.88)	4.61	(.99)
Investigative Interests	C	2.00	(1.43)	3.44	(1.61)	4.61	(1.78)
	D	1.97	(1.40)	3.26	(1.50)	4.16	(1.93)

[a] C = Cannon Crewman

D = Motor Vehicle Operator (Driver)

had a strong, negative impact on these particular results. The demands of the task appeared to impose a unique kind of restriction in the range of possible ratings. That is, the effective range of ratings within levels covered only two or three points rather than the entire seven points. This outcome served to reduce within-subjects variability, as all ratings fell close together. Although SMEs were clearly advised not to respond according to belief that "better must mean more," the mean ratings suggest that a demand characteristic was created by the instructions to rate at three levels. The result was ratings of attribute levels which correspond to level of performance, with ceiling effects occurring at the highest level. These effects would explain the extremely low reliabilities for Physical/Psychomotor attributes at the 85th percentile.

The fact that attribute requirements were elicited for three performance levels also may have clouded the findings in another way and reduced interrater agreement, i.e., increased between-subjects variance. Although definitions were provided for the three performance levels, how the SMEs actually interpreted these definitions was unknown. SMEs may have had different interpretations of the attributes from our definitions as well as from one another. For example, their verbal reports seemed to indicate some tendency to interpret performance levels in terms of particular soldiers in their charge, rather than from a more general (and shared) view of job performance at a particular level. They also tended to rate attributes in terms of the characteristics of someone who performed at that level, rather than in terms of the actual requirements of the job. The performance criterion, then, was more ambiguous than expected, pointing out a clear need for a very specific definition of the criterion. It was apparent that understanding the task requirements -- what was meant by the performance levels and how to do three ratings at a time -- took more time and energy than actually doing the ratings. In short, the use of three performance levels may have made the task harder than was intended, and interfered with the SMEs' ability to rate true requirements.

Two other factors may have contributed to low interrater agreement. SMEs were not given written descriptions of what they were to rate. Instead they were asked to decide individually the nature and content of entry level work and, specifically, what it required in terms of attributes. Moreover, they were to rate the whole job -- all work within all duty positions -- and not just some specific task or set of tasks. This very broad scope allowed considerable opportunity for variance. As a result of personal experiences and/or selective memory, the SMEs could differ a great deal in what they were evaluating. Obviously, higher interrater agreement would be expected for narrower areas of consideration. In addition, some SMEs found the scale anchors frustrating rather than helpful. Raters appeared to have difficulty using anchors as reference points for comparing tasks within their MOS. Some tended to evaluate the job in terms of whether the exact tasks depicted were or were not an actual part of the job. With some anchors that depicted common soldier tasks, some SMEs had problems separating the overall soldier requirements from the specific job requirements. Thus, although very familiar behaviors were thought to be the best for illustrating a level of an attribute, this was not necessarily the case.

The results of the ANOVA indicate that attribute profiles for the two MOS are not significantly different. The effects that were significant, Attributes, Performance Levels, and their interaction, are most likely a function of the high statistical power related to the large number of

degrees of freedom, and so are not really meaningful. Despite this, the data provide some useful information. The minimal differences which do occur suggest that some differences (as well as similarities) between MOS may exist, but may be masked in the present research. In addition, rank orders of the magnitude of ratings were different for both MOS at all performance levels, again suggesting there may be some differences in patterns of attributes which need further examination. For instance, at the 85th percentile, Verbal Ability ranked tenth for Cannon Crewman but third for Motor Vehicle Operator, while Stamina ranked first and fifteenth respectively. If one were to select only the five variables with the highest ratings, the selection would be different for each MOS. However, the top five are not necessarily the most important attributes: They are ranked on level of required attribute and not on relative importance of the attribute.

In summary, NCOs appeared to understand, in general, how to use the set of scales constructed to rate job requirements. The requirement for three sets of ratings simultaneously, however, created some problems. First, the actual physical arrangement of the scales on the page confused people. Second, it seemed to impose limits on the magnitude of ratings assigned. Given the expanse of the criterion to be rated -- the entire MOS at Skill Level 1 -- and the limitations created by the design itself -- different performance levels -- the obtained indices of interrater agreement are reasonable.

These findings suggest that better reliability estimates could be obtained with fewer raters if SMEs were asked to rate requirements for a single performance level; i.e., to estimate the minimum level of an attribute required to perform the job successfully. Further, more reliable ratings may be obtained by changing to a generic set of scale anchors (e.g., very low,low, moderate, etc.) or otherwise replacing the present behavioral anchors and/or focusing raters' attention on evaluating a specific task, a well-defined set of tasks, or a written job description would yield better reliability coefficients. Elimination of the restriction in range of ratings which was created by including three performance levels, should yield better discrimination among the attributes within MOS, and differences in attribute profiles across MOS.

References

Fleishman, E. A. (1982). Systems for describing human tasks. American Psychologist. 30, 1127-1149.

Fleishman, E. A., & Quaintance, M. K (1984). Taxonomies of human performance. Orlando, FL: Academic Press Inc.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87, 1-28.

Rossmeissl, P. G. (1985, October). Computerized approaches for estimating ability requirements. Paper presented at the 26th Annual Conference of the Military Testing Association, San Diego, CA.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984, August). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada.

METHODOLOGICAL PROBLEMS IN IDENTIFYING ABILITY REQUIREMENTS RELATED TO SOLDIER PERFORMANCE

Dr. Claude R. Miller

US ARMY TRADOC
SYSTEMS ANALYSIS ACTIVITY
(TRASANA)

INTRODUCTION

The Training Effectiveness Analysis (TEA) Division of TRASANA recently completed a basic study of the procedures necessary to represent relationships between soldier variables and system performance in computerized combat models (1). The study was prompted by a concern for future system manning requirements. Census data indicate that the number of individuals available for recruiting is declining and will continue to decline well into the next decade (2). At the same time, the Army is undergoing a force modernization program that includes technologically advanced weapon systems that put a premium on highly skilled operators and maintainers. These concurrent processes indicate a need to identify soldier variables related to system performance so that the available manpower may be distributed among the various weapon systems in a manner that maintains or improves combat effectiveness.

Combat models may prove useful for estimating the extent to which relationships between soldier variables and performance affect combat effectiveness. However, prior to modeling, existing relationships between soldier variables and performance must be identified and mathematically described. The term "soldier variables" may refer to any number of physical or psychological attributes of an individual. Soldier variable data normally available, or easily obtained, include physical profiles, demographic descriptors, and results from the Armed Services Vocational Aptitude Battery (ASVAB). At present, the Army assigns soldier to Military Occupational Specialties (MOSs) on the basis of ASVAB scores and manpower needs. The usefulness of ASVAB scores to predict soldier performance in the field remains a matter of study. Supplemental and/or alternative selection and assignment procedures may be required in the future.

One possible supplemental/alternative approach is based on the work of Fleishman (3) who has studied the relationships between human abilities and task performance. Fleishman's work has suggested a taxonomy of cognitive, perceptual, and psychomotor abilities differentialiy related to performance on various types of tasks. It was in the context of Fleishman's work that an attempt was made in the TRASANA study to identify soldier variables related to system performance. This report summarizes some of the difficulties encountered in that effort.

METHOD AND RESULTS

The referenced TRASANA study focused on Air Defense Artillery soldiers in MOS 16S, STINGER gunners. Subsequent to the referenced study, additional data were collected on five other Air Defense Artillery MOSs and two Field Artillery MOSs. Subject Matter Experts (SMEs) were asked to rate the extent to which each of 33 separate basic human abilities, taken from the work of Fleishman, was related to MOS critical task performance.

It was immediately obvious that having SMEs rate abilities required for each MOS critical task would impose an unreasonable workload on the SMEs that could produce results of questionable validity and reliability. For that reason, the survey instructions asked the SMEs to rate each ability in terms of how necessary the ability was to "successful performance in the MOS" without reference to specific critical tasks.

The format of the survey itself included:

1. The name of the ability (eg., Reaction Time)
2. A definition of the ability in the form of a question (eg., Does a gunner have to be able to react quickly to sights or sounds?)
3. A space to check "Yes", "No", or "Do Not Understand the Question."
4. A graphic seven-point rating scale. The scale included the integers from 1 to 7 and graphically indicated midpoints between integers. The only rating guidance provided on the scale itself was entered next to the numbers 1, 4, and 7. Next to the number 1 was the description "Ability Needed But Only To A Small Degree." Next to the number 4 was the description "Ability Needed To A Moderate Degree." Next to the number 7 was the description "Ability Needed To The Highest Possible Degree."

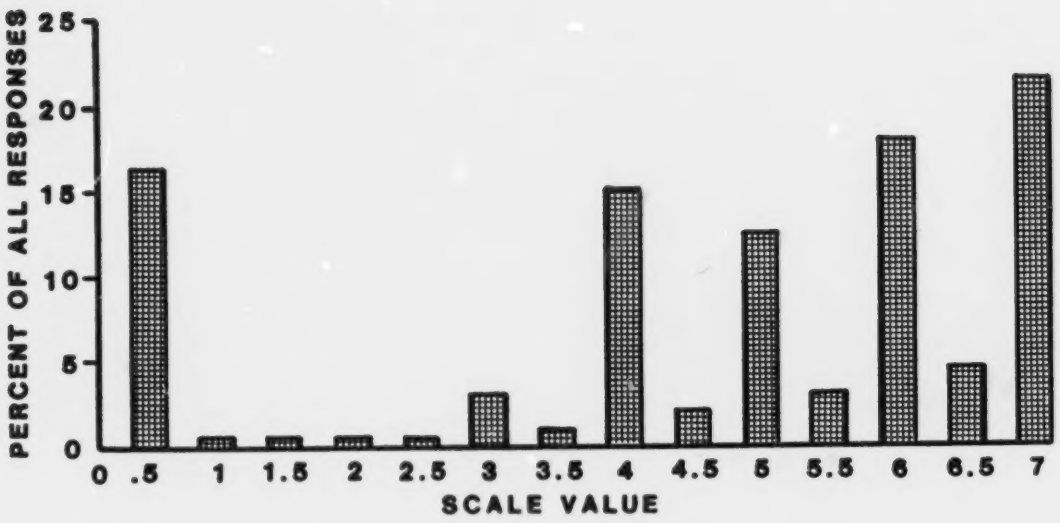
Prior to administration of the survey to SMEs, the survey was "pilot tested" with military personnel at TRASANA. Results from the pilot test indicated that many soldiers had difficulty understanding the question, i.e. the definition of the ability. Consequently, the definitions were rewritten using simpler language. Psychologists in the TEA Division were asked to compare the original and simplified definitions for comparability and generally agreed that the essence of each definition had been preserved. Nevertheless, opinion was not unanimous and some concern remained over whether all simplified definitions were accurate translations of the originals.

The final form of the survey was administered to a total of 85 SMEs from eight different MOSs. The number of SMEs for each MOS varied from n = 6 to n = 16. As stated previously, a seven-point rating scale was used for SMEs to rate how much of each ability was necessary to perform successfully in the MOS. Alternatively, the SMEs could check the space provided indicating that the ability was not needed at all. If an SME checked that an ability was not necessary, that response was given a scale value of zero.

Actual results in terms of how the abilities were rated for each MOS will not be addressed. Instead, the response pattern will be discussed. Given 85 SMEs rating 33 different abilities, a total of 2805 responses were possible. Of these, 219 (7.8%) were omitted, i.e. no scale response and no check that the ability was not needed. The remaining 2586 responses showed a distinct pattern that was very consistent across all MOSs. First, there was little use of the midway points graphically indicated on the scale. Only 12% of all responses were placed on the midway points. The preponderance of responses were given to the scale values 4, 5, 6, and 7. The scale values of 1, 2, and 3 were virtually ignored. The most frequently used scale value was 7 (22.7% of all responses). Figure 1 summarizes the response distribution pattern across the seven-point scale.

One final note regarding the pattern of responses seems of interest. The standard deviations among SME ratings for the highest rated abilities were consistently and significantly lower that the deviations among the lower rated abilities. The result was consistent for each group of SMEs surveyed.

Figure 1. Distribution of SME Responses to the Ability Rating Scales



DISCUSSION

The primary emphasis of the TRASANA study was to represent the relationships between soldier variables and performance in combat models using existing soldier variable data, not to identify a new set of soldier variables to explore. As a result, the relationships suggested by the SME ratings were not examined in detail, and validation of the SME ratings remains to be accomplished using actual performance data and scores from tests that measure the basic abilities.

Nevertheless, the experience gained from administration of the ability rating scales has implications for future scale development and interpretation. A major consideration is the workload put on the raters when ratings are required by specific tasks. To minimize the adverse impact of lengthy ratings, SMEs were asked to rate the abilities in terms of success in the MOS, not specific tasks. Although that procedure saves time, detailed information is lost. To recover at least part of that information, a group discussion was held with the SMEs after all ratings had been completed. From these discussions, it was possible to determine why certain abilities were rated low or high, and how certain abilities related to specific critical tasks.

A second consideration is how the abilities are defined. The SMEs had difficulty understanding the psychological terms used in many of the definitions, and psychologists had some difficulty with simplified definitions. To ensure common understanding of terms is a long standing principle of survey construction, but one too often assumed. Future scale development should take the time to develop sound and understandable definitions.

A final problem is the tendency of raters to judge an ability as needed to a moderate-to-high degree, or not needed at all. That tendency was noted on the first administration of the survey. In subsequent administrations raters were asked to save their highest ratings for those few abilities strongly required in the job. There is no evidence that such verbal instructions had any effect on the rating pattern. On two surveys, raters actually took their pencil and extended the scale out to 10 points, then checked "10." Use of anchor points may have reduced the tendency toward high ratings, but time constraints precluded development of verbal anchors. Another possible approach to the response pattern problem may be through the use of variance as well as mean ratings to establish the criterion used to identify an ability as needed for a job. Raters tend to agree on those abilities strongly related to successful job performance. The exact method for establishing criteria remains to be determined.

REFERENCES

1. Miller, C. R. Modeling Soldier Dimension Variables in the Air Defense Artillery Mission Area. TRASANA TEA-2-85, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, NM, February 1985.

2. "Soldier 90: Demographic Data And Trends," US Army Soldier Support Center, National Capital Region, Ft. Benjamin Harrison, IN, September 1982.

3. Fleishman, E. A. Toward a Taxonomy of Human Performance, American Psychologist, Vol. 30, 1975.

Comparison of Weapon Systems Using Ability Requirements Scales

Jane M. Arabian
U.S. Army Research Institute for
the Behavioral and Social Sciences

The capability and complexity of weapon systems have been increasing. Meanwhile, census data indicate that there will be fewer 18-24 year olds between 1980 and the 1990's. The greater sophistication of military weapons together with the decrease in the size of the potential military applicant pool make it increasingly important not only to design systems that are within physical and cognitive capabilities of the available personnel but also to match correctly the skills and abilities of individuals to the skill and ability requirements of jobs (Shields & Baker, 1981). Inappropriate personnel classification could result in poor human resource management as well as a failure to benefit from the maximum functional capabilities of the weapon system. The best personnel classification system, however, will not be able to overcome system designs that require unavailable personnel attributes.

In order to optimize man-machine system performance, the engineering design process needs to be sensitive and responsive to the impact of the system on manpower, personnel, and training (MPT) issues. The MPT community, in turn, must provide information to the designers as early as possible in the system development process. The earlier in the process, the more likely it is that information will be based on subject matter experts' (SME) best estimates derived from their professional experience. Later on in the process, information can be based on empirical data obtained from the actual system. However, many of the design decisions which affect MPT occur before it would be possible to collect data with the actual system.

The Army Research Institute conducted a project to examine the Weapon System Acquisition Process (WSAP) and identify critical points in the process for MPT information (Promisel, et al, 1985). The project has been referred to as the Reverse Engineering Project because it traced the development and acquisition process in reverse, from the end-product back to the beginning (Concept Exploration Phase in the WSAP). Scientists working on the project studied as much available documentation, both historic and current, as possible for selected weapon systems.

The goal of the project was to examine how the performance of systems was affected by the specific system requirements and engineering and management decisions made earlier in the development process. For example, with regard to personnel requirements, project staff identified personnel requirements for the system as stated in the acquisition documents (e.g., where did the requirement come from and what kinds of data supported the requirement). Project staff also identified points in the acquisition process where it seemed feasible that more or different kinds of information could or should have been considered. Positive impact on system development from timely personnel information use was also noted. Recommendations were then made about how the acquisition process could be augmented to ensure that personnel requirements information would be developed and applied as early as possible in the WSAP.

The WSAP for Stinger, a man-portable air defense system, and its predecessor, Redeye, were examined first. One of the more salient findings was that very little personnel information was available throughout the acquisition cycle. Requirements for height and ASVAB scores were available, but there was little, if any, information about perceptual and psychomotor skill

and ability requirements. As the project staff became more familiar with the two systems through the requirements documentation, it seemed that the newer system would place more demands on the soldier than the older system. This seemed especially true if the new system's capabilities were to be fully realized. The new system did not appear to take the potentially increased psychomotor requirements into account; soldiers who operated the old system were expected to operate the new system effectively.

The research study described in this report was designed to provide information about the skill and ability requirements of the two weapon systems and to determine whether empirical support could be obtained for the impression that Stinger required more and/or higher levels of specifiable skills and abilities than Redeye. Since this type of information would be particularly valuable early in the WSAP, before the system would be physically available, the feasibility of using rating scales based on expert judgements to collect this information was examined.

METHOD

Subjects. Two separate samples of subjects served as SME raters. The first sample (Scientist Sample) consisted of eight Army Research Institute research scientists who were participating in the Reverse Engineering Project. They were familiar with the WSAP requirements documentation and testing results for the Redeye and Stinger systems. The second sample of SME raters (Instructor Sample) consisted of ten military instructors for the Redeye and Stinger training courses. They had been experienced system operators before becoming training instructors.

Rating Instruments. Both samples of raters employed the Fleishman abilities taxonomy (Fleishman, 1975). The list of abilities appears in Table 1.

Table 1

Fleishman Ability Taxonomy

1.	Oral Comprehension	21.	Visualization
2.	Written Comprehension	22.	Static Strength
3.	Oral Expression	23.	Explosive Strength
4.	Written Expression	24.	Dynamic Strength
5.	Memorization	25.	Trunk Strength
6.	Problem Sensitivity	26.	Stamina
7.	Originality	27.	Extent Flexibility
8.	Inductive Reasoning	28.	Dynamic Flexibility
9.	Category Flexibility	29.	Gross Body Equilibrium
10.	Deductive Reasoning	30.	Speed of Limb Movement
11.	Information Ordering	31.	Gross Body Coordination
12.	Math Reasoning	32.	Multi-Limb Coordination
13.	Number Facility	33.	Wrist-Finger Speed
14.	Fluency of Ideas	34.	Finger Dexterity
15.	Time sharing	35.	Manual Dexterity
16.	Flexibility of Closure	36.	Arm/Hand Steadiness
17.	Speed of Closure	37.	Control Precision
18.	Selective Attention	38.	Rate Control
19.	Perceptual Speed	39.	Reaction Time
20.	Spatial Orientation	40.	Choice Reaction Time

An ability rating scale for each ability includes a definition of the ability and a rating scale anchored with behavioral examples. A rating of "1" indicated that the rater estimated that very little of a particular ability would be needed for a job or task; a rating of "7" indicated that a great deal of the ability was estimated as necessary. In addition to the scales themselves, binary decision flow diagrams have been developed (Malamad, et al, 1980). The flow chart leads the rater through a series of yes/no decisions which, when completed, results in specification of the abilities required to perform the job. Figures 1 and 2 provide examples of

the flow diagram and ability scales, respectively. The scale on the left uses military examples for anchors, as in the computerized assessment system described below; the scale on the right in Figure 2 uses non-military examples, as in the paper and pencil version of the ability assessment system.

Figure 1

Flow Diagram

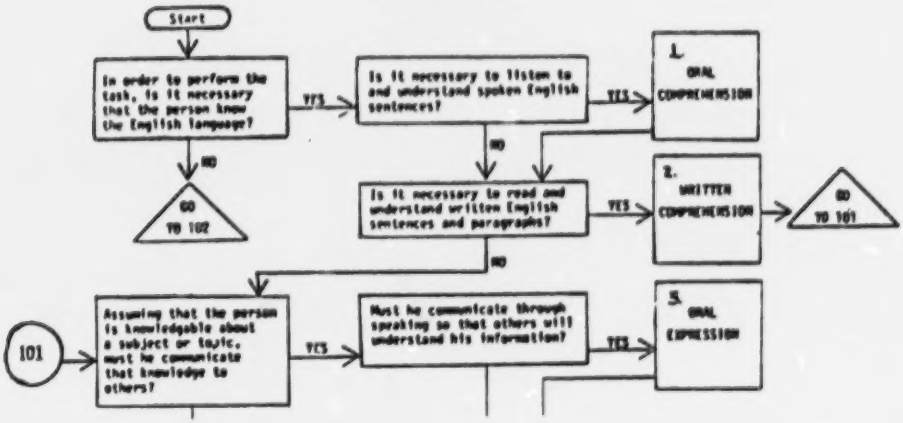
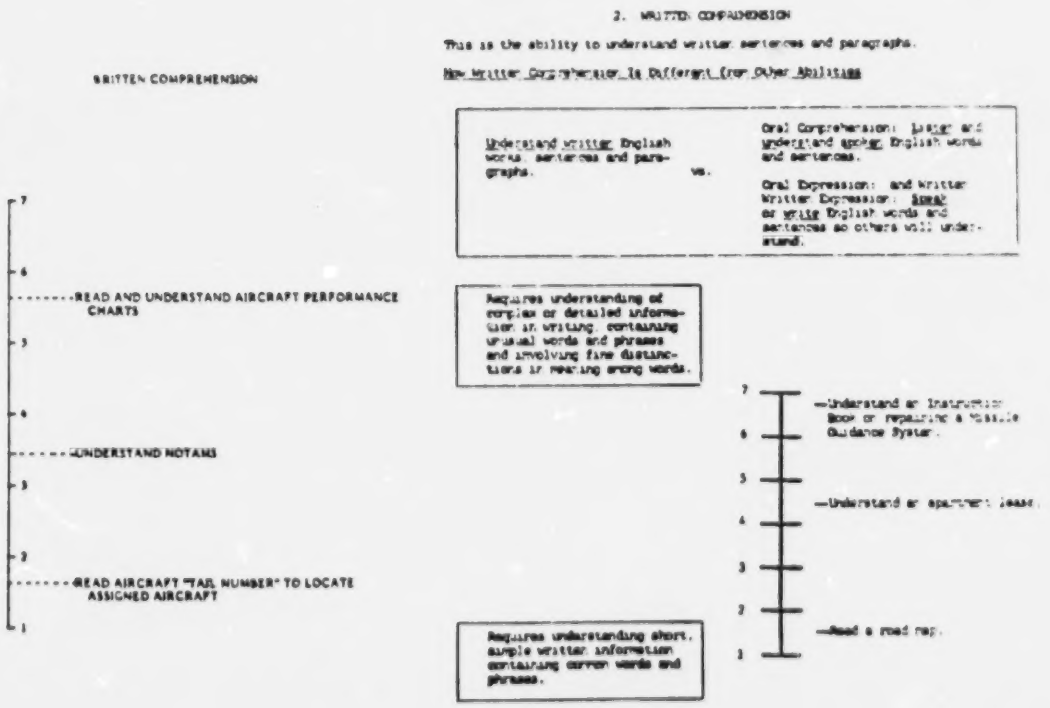

Start
In order to perform the task, is it necessary that the person know the English language?
YES
Is it necessary to listen to and understand spoken English sentences?
YES
1. ORAL COMPREHENSION
NO
GO TO 102
Is it necessary to read and understand written English sentences and paragraphs?
YES
2. WRITTEN COMPREHENSION
GO TO 101
NO
101
Assuming that the person is knowledgable about a subject or topic, must he communicate that knowledge to others?
YES
Must he communicate through speaking so that others will understand his information?
YES
3. ORAL EXPRESSION

Figure 2

Ability Scales


WRITTEN COMPREHENSION
7
6
READ AND UNDERSTAND AIRCRAFT PERFORMANCE CHARTS
5
4
UNDERSTAND NOTAMS
3
2
READ AIRCRAFT "TAIL NUMBER" TO LOCATE ASSIGNED AIRCRAFT
1
2. WRITTEN COMPREHENSION
This is the ability to understand written sentences and paragraphs.
How Written Comprehension Is Different from Other Abilities
Understand written English words, sentences and paragraphs. vs.
Oral Comprehension: Hear and understand spoken English words and sentences.
Oral Expression: and Written Written Expression: Speak or write English words and sentences so others will understand.
Requires understanding of complex or detailed information in writing, containing unusual words and phrases and involving fine distinctions in meaning among words.
7
6
—Understand an instruction book on repairing a missile Guidance System.
5
4
—Understand an apartment lease.
3
2
—Read a road map.
1
Requires understanding short, simple written information containing common words and phrases.

Raters in the Scientist Sample employed a computerized version of the flow diagrams with ability scales (Rossmeissl, et al, 1983). This computerized ability assessment program was written in Applesoft© BASIC language and presented to the raters on an APPLE II microcomputer.

Raters in the Instructor Sample used a paper and pencil format to perform the ability assessments. The decision flow diagram was presented in one booklet with the ability scales in a separate packet. Raters worked through the flow diagram and when they came to a point for rating the level required of an ability they switched to the packet of scales, located the correct rating scale, made their rating, and returned to the flow diagram to continue the procedure. The scale anchors for the paper and pencil assessment were generic (non-military) behavioral examples (Theologus, et al, 1970).

Procedure. Ability ratings were obtained from each sample of raters separately. The computerized ratings were obtained from one subject at a time. The paper and pencil ratings were collected in one session with all raters present, but discussion of ratings was not allowed until all the assessments had been completed. For both samples, half the raters estimated ability requirements first for Redeye and then for Stinger; the order was reversed for the remaining raters. The raters in the Scientist Sample completed their estimates on two successive days (i.e., one system on Day 1, the other system on Day 2). The raters in the Instructor Sample completed their rating for one system, took a fifteen minute break, and completed their ratings for the second system.

Instructions for both groups were the same. All raters were asked to think about the soldier abilities that would be required during the successful operation of the system (Redeye or Stinger) under combat conditions. The decision flow diagrams and procedure for using the ability scales were explained. The raters were cautioned that the scales' anchor examples had not been validated for Redeye or Stinger and should therefore be used only as a rough guide. The raters were reminded to answer the decision flow questions based on their knowledge and/or experience with the weapon systems, to think about only one system at a time, and avoid making mental comparisons of the two systems.

Analyses. Given the relatively small sample sizes and exploratory nature of this research study, statistical analyses of variance were not performed on the entire data set from either sample. However, one-tailed correlated t-tests were performed on ratings for selected abilities if all raters in a sample agreed that the ability was required for both systems. Statistical analyses comparing the ratings of the two samples were not conducted since the background of the raters (scientists vs. instructors) was confounded with the rating mode (computer vs. paper and pencil).

RESULTS AND DISCUSSION

The number of ratings and range of non-zero ratings for each ability and both systems are presented in Table 2. Sixteen abilities were rated for both Redeye and Stinger by at least half the raters in the Scientist Sample. Five abilities were not rated for either system by any of the scientist-raters. There was unanimous agreement in the Scientist Sample that seven of the abilities were required by both weapon systems. In no case was an ability consistently rated as required for one system but not for the other system.

The raters in the Instructor Sample tended to rate more abilities: every ability was seen as required for both systems by at least half of the raters. There was unanimous instructor-rater agreement that six of the abilities were required by both Redeye and Stinger.

Table 2

Ranges of Ability Ratings

Ability	Scientist Sample (N=8)				Instructor Sample (N=10)			
	REDEYE		STINGER		REDEYE		STINGER	
	n	Range of Non-zero ratings	n	Range of Non-zero ratings	n	Range of Non-zero ratings	n	Range of Non-zero ratings
1. Oral Comprehension	5	5.0-6.0	5	5.0-6.3	9	3.0-6.7	9	4.3-6.7
2. Written Comprehension	4	3.0-5.0	5	3.0-5.0	10	3.3-6.7	9	3.3-6.7
3. Oral Expression	3	5.0-5.7	3	3.0-5.0	10	3.0-6.7	9	4.0-6.7
4. Written Expression	0		0		10	3.0-6.7	9	4.3-6.7
5. Memorization	8	5.0-7.0	8	4.0-7.0	10	2.3-6.7	10	3.7-6.7
6. Problem Sensitivity	8	4.0-5.7	8	4.0-7.0	10	2.0-6.7	10	4.0-7.0
7. Originality	5	3.0-4.7	7	3.0-5.7	10	1.7-5.7	9	1.3-5.3
8. Inductive Reasoning	1	3.3	3	3.0-4.7	10	1.3-6.7	9	2.3-6.7
9. Category Flexibility	3	3.0-4.7	3	2.0-6.0	10	2.0-6.7	9	3.0-6.7
10. Deductive Reasoning	7	3.7-5.7	8	3.7-6.0	10	2.7-6.7	9	3.0-6.7
11. Information Ordering	5	3.0-5.0	3	3.0-6.0	10	2.0-5.7	8	3.7-6.7
12. Math Reasoning	2	1.0-5.3	1	6.0	9	2.3-6.3	8	1.7-6.7
13. Number Facility	3	1.0-5.0	5	1.0-5.0	9	1.3-6.3	7	2.3-6.7
14. Fluency of Ideas	1	2.3	1	3.3	10	1.3-6.3	6	3.3-6.3
15. Time Sharing	4	3.0-6.7	6	1.0-6.3	10	3.3-6.7	10	2.3-6.7
16. Flexibility of Closure	4	5.0-6.0	6	1.7-7.0	10	3.3-6.3	8	3.7-6.7
17. Speed of Closure	1	5.3	0		8	2.7-6.7	8	4.3-6.7
18. Selective Attention	8	2.7-7.0	8	3.0-5.7	9	3.7-6.3	9	3.7-6.7
19. Perceptual Speed	8	3.3-6.0	8	4.0-6.7	10	2.3-6.7	9	2.7-7.0
20. Spatial Orientation	8	4.0-7.0	8	5.0-7.0	10	3.3-7.0	10	4.0-7.0
21. Visualization	4	2.7-6.0	4	3.7-6.0	10	2.3-6.7	9	3.3-6.7
22. Static Strength	2	2.7-4.7	1	5.3	8	3.3-6.7	7	3.7-6.7
23. Explosive Strength	0		0		8	2.7-6.7	6	5.0-7.0
24. Dynamic Strength	0		0		8	2.3-6.3	6	1.3-6.3
25. Trunk Strength	0		0		7	2.7-6.3	6	1.7-6.7
26. Stamina	4	3.0-5.0	4	4.0-5.0	9	3.3-7.0	10	1.7-6.3
27. Extent Flexibility	5	2.7-6.0	5	2.3-6.0	9	4.0-6.3	10	4.0-6.7
28. Dynamic Flexibility	0		0		9	3.3-6.7	9	2.3-6.3
29. Gross Body Equilibrium	7	2.3-3.3	6	2.0-3.3	9	3.3-6.3	9	3.3-6.7
30. Speed of Limb Movement	5	3.0-5.0	6	2.0-5.3	10	3.7-6.7	9	4.3-6.7
31. Gross Body Coordination	1	4.0	1	4.0	9	3.7-6.7	9	2.3-6.7
32. Multi-Limb Coordination	6	1.0-7.0	4	3.0-5.0	8	3.3-6.3	7	2.3-6.7
33. Wrist/Finger Speed	3	3.0-5.3	1	4.0	10	2.7-6.7	10	2.7-6.7
34. Finger Dexterity	3	3.0-5.0	1	4.0	10	3.0-6.3	9	2.3-6.7
35. Manual Dexterity	3	3.0-5.7	0		10	2.7-6.3	9	3.3-6.7
36. Arm/Hand Steadiness	8	2.0-5.7	8	2.7-6.0	10	4.3-6.7	10	3.0-6.7
37. Control Precision	1	5.3	3	5.0-6.7	9	2.3-6.3	9	2.7-6.3
38. Rate Control	1	3.3	0		9	2.3-6.7	10	3.0-6.3
39. Reaction Time	8	3.0-7.0	8	4.0-7.0	8	4.7-7.0	7	4.3-6.7
40. Choice Reaction Time	0		0		8	2.7-6.7	8	3.3-6.7

The mean ability level ratings for abilities rated as required for both systems by all raters in a sample are presented in Table 3. The means pro-

TABLE 3

Mean Ratings for Abilities
Unanimously Rated as Required

Ability	Scientist Sample			Instructor Sample		
	Redeye $\bar{X}$	Stinger $\bar{X}$	t p<.05	Redeye $\bar{X}$	Stinger $\bar{X}$	t p<.05
5. Memorization	6.3	6.3	n.s	5.5	5.5	n.s
6. Problem Sensitivity	5.0	5.3	n.s	4.8	5.5	n.s
15. Time Sharing	(4.7)a/	(3.8)	n.s	5.3	4.5	2.26
18. Selective Attention	4.4	4.7	n.s	(5.0)	(5.1)	2.16
19. Perceptual Speed	4.4	5.2	n.s	(5.1)	(4.8)	n.s
20. Spatial Orientation	5.8	5.9	n.s	5.5	5.6	n.s
33. Wrist/Finger Speed	( )	( )		5.1	5.0	n.s
36. Arm/Hand Steadiness	4.0	4.3	n.s	5.6	5.4	n.s
39. Reaction Time	5.6	5.7	n.s	(4.7)	(4.0)	n.s

a/ Numbers in parenthesis are based on fewer than 8 ratings for the Scientist Sample or 10 ratings for the Instructor Sample.

vide some evidence that Stinger may place more demands on the abilities of soldiers than Redeye, as indicated by higher ability level ratings for some abilities. However, the mean differences are not overwhelming and generally do not approach statistical significance. Nevertheless, the data do suggest that Stinger is certainly not less demanding than Redeye.

Although the data do not fully support the notion that Stinger requires more or higher levels of abilities that Redeye, the mean ratings do help to identify "high drivers" such as memorization, problem sensitivity, and spa-

tial orientation. Relatively high levels of these three abilities are required by both systems. Had this type of information been available early in the system design process, it may have been possible to modify the design of the newer system to reduce the demand for these abilities.

The potential usefulness of ability requirements ratings are limited, however, by the lack of empirical data. The relationship between some estimated level of an ability requirement and actual job or task performance needs to be determined. Without this information it is not possible to establish what effect an individual with a "5" level, as opposed to a "6" or "4" level, of an ability would have on system performance if the requirement was estimated to be at level "6" for that ability. Information about the distribution of abilities in the population is also needed. If most individuals in a given population possess high levels of certain abilities then it may not be necessary to design systems that require low levels of those abilities. On the other hand, if few individuals have high levels of particular abilities, it may be wise to avoid system designs that require high levels of those abilities.

Whether or not an ability rating methodology can be developed into a sensitive psychometric instrument, the rating approach used in the research study presented in this paper appears to have potential applications in the design process. At the least, a procedure which can identify "high driver" abilities before a system has been fully designed and built may be helpful to design engineers and allow them to avoid or modify designs that are likely to be too demanding on the psychomotor and perceptual capabilities of personnel.

ACKNOWLEDGMENTS

This research was conducted while the author was a member of the Task Force for the Reverse Engineering Project. The opinions, views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, expressed or implied, of the U.S. Army Research Institute for the Behavioral and Social Sciences or the Department of Defense or the United States Government.

REFERENCES

Fleishman, E.A. (1975). Toward a taxonomy of human performance. American Psychologist, 30, 1127-1149.

Mallamad, S.M., Levine, J.M., & Fleishman, E.A. (1980). Identifying ability requirements by decision flow diagrams. Human Factors, 22, 57-68.

Promisel, D.M., Hartel, C.R., Kaplan, J.D., Marcus, A., & Wittenburg, J.A. (1985). Reverse engineering: Human factors, manpower, personnel, and training in the weapon systems acquisition process (ARI Technical Report 659). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rossmeissl, P.G., Tillman, B.W., Rigg, K.E., & Best, P.R. (1983). Job assessment software system (JASS) for analysis of weapon systems personnel requirements (ARI Research Report 1355). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Shields, J.L. & Baker, J.D. (1981). The Army's personnel problems and the golden spike solution. Paper presented at the National Security Industrial Association First Annual Conference on Personnel and Training Factors in System Effectiveness, San Diego, CA.

Theologus, G.C., Romashko, T., & Fleishman, E.A. (1970). Development of a taxonomy of human performance: A feasibility study of ability dimensions for classifying human tasks (AIR Technical Report 726-5). Washington, DC: American Institutes for Research.

Computerized Approaches for Estimating Ability Requirements

Paul G. Rossmeissl
U.S. Army Research Institute

The past few years have witnessed a dramatic increase in the power, sophistication, and availability of microcomputers. It seems natural that this new technology be devoted to addressing the recurring problem of determining what human abilities are needed to do what must be done in a given job. This paper discusses some of the ways in which the new technology could be brought to bear on the problem at hand.

One way to use the computer to investigate ability requirements is to take a tried and true method (one that works without a computer) and to computerize it. That is, to display the instrument on a CRT and store the responses on disc rather than using paper and pencil for these functions. The first part of this paper will review an investigation that followed this approach to the issue.

An alternative technique for using computers in this area is to use the machine to administer methods that are either difficult or impossible to administer in any other manner. The final portion of this paper will propose and discuss a few methods of this type.

Paper-and-pencil vs. Computerized Rating Scales

One of the most commonly used techniques in the determination of ability requirements is the use of ability rating scales (i.e. Fleishman, 1982). Such scales typically include the definition of a human ability and a rating scale anchored with behavioral examples. While the scales are typically presented on paper, they can be easily presented on a computer's CRT as well.

Rossmeissl, Kostyla, and Tillman (1983) conducted an investigation of computerized versus paper-and-pencil modes of rating scale presentation. They used rating scales of 30 human abilities developed by Rossmeissl and Dohme (1982). These rating scales are very similar to those developed by Fleishman (1975, 1982), but contain anchor stimuli that are directly relevant to jobs within Army aviation. The computerized presentation of the scales was implemented by Rossmeissl, Rigg, and Best (1983) using the binary decision flow diagrams developed by Mallamad, Levine, and Fleishman (1980). The two procedures were evaluated by having experienced Army aviators rate the abilities required to perform an Army aeroscout helicopter mission. Half of the aviators estimated ability requirements using paper-and-pencil rating scales, while the other half used rating scales presented on an Apple II microcomputer. All of the ratings were made along a seven point scale.

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

The findings of this research showed that the two methods gave very similar results. Both methods showed very high inter-rater agreement. The inter-class correlation coefficient (ICC) for the paper-and-pencil scales was .91, while the ICC for the computer presented scales was .96. Both methods were also able to produce ratings that discriminated among the various aptitude requirements (both F values > 190 and significant at the .001 level).

This result is perhaps the most important aspect of the research, in that despite the somewhat popular belief of everything being better when its on a computer, the computerized procedure did not show superior psychometric properties in comparison to the traditional paper-and-pencil forms. On further thought, however, this finding really should not be surprising, since after all the same rating scales and much the same definitions were used in both cases.

The only real difference in results between the two procedures was that the computerized method led to ratings of not required for four abilities that were rated as required with the paper-and-pencil forms. These four abilities are given in Table 1. Subsequent analyses revealed that the discrepancy in ratings given to reaction time was an artifact of the

Table 1
Abilities Rated as Not Required with the Computer
but as Required with Paper-and-Pencil Forms

Ability	Mean rating with paper-and pencil forms
Stamina	6.1
Closure flexibility	5.4
Reaction time	5.3
Static Strength	4.8

decision diagrams used to program the computer. The reason behind the other differences was not perfectly clear, but was most likely attributable to subtle differences between the two methods in the ability definitions. Such small differences could easily be rectified with future versions of the instruments.

Given that the two procedures produce similar results in a measurement or psychometric sense, a choice between the two instruments would depend upon other considerations. Each procedure has advantages for particular situations. The paper-and-pencil forms are inexpensive and can be administered in groups or by mail. These considerations are especially compelling when the instrument needs to be administered to a large number of job experts to get the required data.

The computerized procedure, however, does have advantages when the subject matter experts are in short supply or have limited availability

for testing. One of these advantages is time. Rossmeissl et al. noted that the mission requirements could be evaluated in about fifteen to thirty minutes using the computer, while it would take from thirty to forty-five minutes to complete the task with the paper-and-pencil rating scales. Also, a debriefing of the participants in this research showed that those who rated ability requirements with the computer found the task more enjoyable than those who used the paper-and-pencil scales. Several participants in the computer group offered to perform the task again, perhaps for some other mission. Such a desire seemed to be far from the mind of anyone who had used the paper-and-pencil scales.

Other Computerized Procedures

The approach outlined above where one simply computerizes an already popular paradigm makes minimal use of the power inherent in the microcomputer. An alternative approach is to develop ability requirement assessment procedures that are designed especially for computer administration. The remainder of this paper will discuss some of the procedures of this sort currently be considered by the Army Research Institute (ARI).

One of these approaches would maintain the use of rating scales to estimate the ability requirements, but use Artificial Intelligence (AI) techniques to provide a more powerful and easy to use tool. Current hardware and software capabilities make it possible to develop a system that "learns" from the user as well as one that allows the present user to learn from the responses of previous users. Such a system would select questions and negotiate responses with the user using the responses of the previous users as the basis of negotiation. Rossmeissl, Rigg, and Best (1983) presented an outline of how such a system might work.

The system would be menu driven to elicit information about the boundary conditions of the job being assessed. The highest order menu might contain global job titles like mechanic, driver, instructor, pilot, etc. and ask the user to select the job most like the one being assessed. If the user selected "mechanic," the next menu would present more specific information about mechanics such as automotive mechanic or tracked vehicle mechanic, and ask the user to select that job closest to the job being assessed. The final menu in this sequence might ask the user to select that job closely resembling the one being analyzed. For example if the user had selected tracked vehicle mechanic, he or she would be presented with a menu including jobs or systems in this area, like M113 armored personnel carrier or M1 Main Battle Tank.

A selection from the final menu would initialize scales and data matrices with the boundary conditions for a job of mechanic in that particular system. The user would then proceed through the branching scales "teaching" the software how the new job differs from the one within the reference system. The user would still branch around inappropriate scales as in the earlier procedure, but if he or she skipped a scale which had been a requirement in the reference job, the software would switch into a negotiation mode, much like the Delphi procedure, and seek to determine what accounts for the difference in job ratings. Such a procedure would also occur when the user rated a scale especially high or low in comparison with that scales rating in the reference job or system.

The key aspect in programming a system of this type is the use of AI procedures to control the presentation the material. In this case the software would be constructed within the framework of a work breakdown structure in which the individual functions, procedures programs, and utility programs are all specified in a formative front end analysis. Within this analysis the individual blocks of the system each perform a single, non-redundant function and are controlled by an executive program. As the system is used the executive program adds this information to the data base containing the boundary or reference job, and in this continue to reduce uncertainty about the new job as well as jobs that had been previously entered. A system of this sort would make better use of the power of the computer, while maintaining the traditional rating scale approach to estimating ability requirements.

As noted previously, such a system is well within the capabilities of current hardware and software systems and a prototype could be developed quite easily. The problem in producing an operational system is that a great deal of basic psychological research needs to be accomplished before such a system could be meaningful. Such work would include: determining the proper number and type of boundary conditions at each menu level, scaling the reference jobs or systems, and developing the appropriate type of scale for the final assessment stage. This last step is particularly difficult in that this scale should relate to the reference system but still allow for comparisons across all possible jobs that might be analyzed.

The speed which computers can present and store data has also led ARI to consider ability requirement analysis procedures that are not dependent on rating scales. One such procedure makes use of Thurstone's law of comparative judgement. Thurstonian scaling uses pair-wise presentation of the data to produce an interval scale for the stimuli. Scale values obtained in this manner are directly comparable across both jobs and abilities. Besides being able to make use of the rich statistical theory underlying the law of comparative judgement, such a method has advantages over rating scales in that one need not be concerned with the development of appropriate anchor stimuli and that the procedure is easier to understand and perform than anchored rating scales.

The disadvantage of the traditional Thurstonian approach is that it is that it can be time consuming to both administer and to analyze. For example, to obtain scale values for say twenty abilities would require $(20!-2!)/2!*(20-2)! = 190$ comparisons. If this procedure were executed on paper forms or cards, not only would the comparisons themselves take considerable time, but the raw data would require tedious hand analyses or keypunching into a computer.

Computerizing this process would clearly simplify the job of analysis, but it could also conceivably reduce the number of comparisons required to scale the abilities. This result could be accomplished by having the computer first present all of the abilities asking the user which are required for the job. All abilities estimated as not required could be assigned a scale value of zero. The computer could then arrange the abilities that were thought to be required in pairs for scaling. If

the twenty abilities mentioned above could be reduced to say fifteen the number of pair-wise comparisons required to complete the task could be reduced to (15!-2!)/2!*(15-2)! = 105 or 55% of the comparisons required with the full ability set. In such a case the computer is doing what it does best: inputing information in one form and quickly processing and converting it into another form.

Conclusion

This paper has outlined a few ways that computer technology can be brought to bear on the problem of estimating human ability requirements within a particular job. An important feature of any such system is that the technology should a be means in arriving at a reliable, valid, and useful analysis tool and not a goal in itself. It is sometimes easy to view a technologically advanced instrument as impressive just because of its technology. The final judgement of any instrument should be how well it does the job at hand, not how it does it. In this case, it is important to remember that a good computerized ability analysis package must be firmly based on empirical research and data before it can be truly useful.

References

Fleishman, E. A. (1975), Toward a taxonomy of human performance. American Psychologist, 30, 1127-1149.

Fleishman, E. A. (1982), Systems for describing human tasks. American Psychologist, 37, 821-934.

Mallamad, S. M., Levine, J. M., and Fleishman, E. A. (1980), Identifying ability requirements by decision flow diagrams. Human Factors, 22 57-68.

Rossmeissl, P. G., and Dohme, J. A. (1982), Using rating scales to determine the aptitude requirements of Army systems. Proceedings of the 24th Annual Conference of the Military Testing Association, San Antonio, TX.

Rossmeissl, P. G., Kostyla, S. J., and Tillman, B. W. (1983), Initial Test and Evaluation of a Computerized Ability Assessment Technique. Paper presented at the meetings of the American Psychological Association, Annaheim, CA.

Rossmeissl, P. G., Rigg, K. E. and Best, P. R. (1983), Job Assessment System (JASS) for Analysis of Weapon Systems Personnel Requirements. U. S. Army Research Institute for the Behavioral and Social Sciences Research Report # 1355, Alexandria, VA.

Thurstone, L. L. (1927), A law of comparative judgement. Psychological Review, 34, 273-286.

Chaparral Crew Performance in the Realistic Air Defense Engagement System

Sarli, Gary G., Johnson, David M., and Lockhart, John M.
United States Army Research Institute
for the Behavioral and Social Sciences
Fort Bliss Field Unit, Fort Bliss, Texas

One of the dangers U.S. forces may face in a future conflict is the attack helicopter. These helicopters are capable of approaching their targets at an extremely low altitude, and need unmask (rise) only briefly — 20 seconds or less — in order to engage their targets (Personal communication, Directorate of Combat Developments, April 1985. The Chaparral is a member of the Army's family of Short Range Air Defense (SHORAD) weapon systems which defends against low-flying fixed-wing aircraft (jets) and rotary-wing aircraft (helicopters) at ranges of less than ten kilometers. Chaparral is vehicle-mounted and fires infrared-seeking missiles; the other SHORAD weapons are the man-portable Redeye and Stinger missile systems, and the Vulcan vehicle-mounted gun system (FM 44-1, 1976). All SHORAD weapons employ at least two soldiers: A chief or leader, and a gunner. The leader is responsible for receiving alerts, visually identifying aircraft, and ordering the gunner to fire. The gunner is responsible for activating the weapon, visually acquiring and tracking aircraft, and firing the weapon. Other crew members (in the case of Chaparral, the driver and assistant gunner) help emplace the weapon and then search for aircraft. Actions taken upon detection of an aircraft will vary depending upon many factors, including the aircraft's specific identity and its actions (FM 44-3, 1977; FM 44-6, 1980).

It is vital that the team or crew make the correct engagement decision in a timely manner to avoid destroying friendly aircraft or allowing hostile aircraft to destroy friendly targets. Inasmuch as CONUS Chaparral crewmembers have been shown to misidentify aircraft 18% of the time (TEA 12-81) the Improved Chaparral has been equipped with an Identify Friend or Foe (IFF) electronic interrogation system. When activated by the gunner, the IFF system sends a coded message to the aircraft and classifies the aircraft as a "true friend", "possible friend", or "unknown", depending upon the aircraft's reply. The IFF classification tones are heard by the senior gunner and the squad leader over their headsets. Of course, the IFF system is not perfect; aircraft can be misidentified due to operator error or system damage.

In order to afford friendly aircraft maximum protection, SHORAD crews and teams operate under weapon control statuses. There are three weapon control statuses: weapons hold, tight, and free. In weapons hold, the gunner only fires if attacked. In weapons tight, if an aircraft is identified by IFF or visually as a true friend, it is tracked until it leaves the area. If the IFF classifies the aircraft as a possible friend or as an unknown, the aircraft must be visually identified as hostile before it can be engaged. In weapons free status, true and possible friends are treated the same as in weapons tight; however, if the IFF classifies the aircraft as unknown, it may be engaged by Chaparral unless local directives and/or SOPs say otherwise. Of course, the gunner may always fire in defense of himself or his defended asset (FM 44-4, 1984). It can be seen that the IFF subsystem and the weapons control status complicate the engagement process by providing more data to be processed within the same amount of time and requiring differing responses depending upon the weapons control status.

The present experiment examined Chaparral crew engagements of scale model helicopters under various realistic conditions in an effort to answer questions such as: Is engagement performance different as a function of weapons control status? Does aircraft intent interact with the effects of weapons control status? How does IFF return interact with weapons control status?

The independent variables were 1) Aircraft intent (friendly or hostile), 2) weapon control status (tight or free), and 3) IFF return (possible friend or unknown for hostiles; true friend or unknown for friends).

The dependent variables were total engagement time; times at detection, target handoff (the process by which the squad leader helps the gunner to find the target), visual acquisition, identification, weapon lock-on, and launch; numbers of correct and incorrect target handoffs, identifications, and engagement decisions; and whether or not the missile hit the aircraft. Summary performance measures included the percentage of correct identifications, the percentages of hostiles and friends which were engaged, the percentages of all hostiles and all friends which were killed, the percentages of engaged hostiles and friends which were killed, and the percentage of hostile targets releasing ordnance.

Method

Subjects

A total of eight Chaparral crews stationed at Fort Bliss, Texas, served as subjects. The Chaparral may be manned by a crew of four or five soldiers; four-man crews were used in this experiment. Each crew contained the following personnel, all of whom have a Military Occupational Specialty (MOS) Code of 16P: Squad leader (rank E5 or E6); Senior gunner (E4 or E5); Assistant gunner (E2 - E4); and Driver (E2 - E4).

RADES Equipment

The Realistic Air Defense Engagement System (RADES) uses realistic scale model aircraft which perform realistic maneuvers in an outdoor environment against actual air defense weapons. Live rounds are not fired; rather, a computer determines what the rounds' flight path would have been, and whether or not the aircraft would have been hit. The RADES system is located on a 2 x 2 kilometer site at Condron Field, White Sands Missile Range, New Mexico. This area, located in the Chihuahuan Desert, contains vegetated sand dunes up to about two meters high in the test area, and mountains approximately 10 kilometers to the west and 60 kilometers to the south. Visibility is usually in excess of 60 kilometers.

The portion of the RADES system used in the present experiment consisted of scale model aircraft, a weapon-computer interface, and computers which were used for scenario presentation and data storage. The RADES target aircraft were non-flying 1/5 scale model Soviet Hind-D and U.S. Cobra helicopters. The helicopters were mounted upon three stands which could be raised and lowered five meters under computer or manual control. These stands were hidden behind sand dunes so that the helicopters could not be seen by the weapon crews when the stands were lowered. In this experiment, the stands were positioned 300 to 450 actual meters (1500 to 2250 scale meters) from the Chaparral crew and always faced at least 20 degrees away from the crew. All the helicopters were equipped with a representative infrared

radiation (IR) source so the Chaparral IR seeker would have a realistic probability of locking onto them. To provide feedback to the crews, each helicopter was also equipped with a device which emitted smoke for several seconds if the helicopter was "destroyed".

A computer at RADES HQ obtained information such as azimuth, elevation, and firing from an electronic interface connected to the Chaparral; information such as time of detection was obtained from the crews' communications network. The computer then determined times for each critical engagement event and whether or not the engagement was successful. Another computer at RADES HQ stored each engagement scenario and the data produced during the running of that scenario.

Military Equipment

Military equipment which was used included one Chaparral Missile System equipped with the IR training round MIM-72F guidance seeker (which simulates the seeker on the MIM-72C improved missile), mounted upon an inert M30 training missile. One IFF subsystem training set, three 7x50 binoculars, two TA312 field telephones with wire, and one 24N tool kit were also used.

During engagements the senior gunner sat in the gunner's compartment of the Chaparral, the driver occupied an observation post 20 meters south of the launcher unit, and the squad leader and the assistant gunner occupied a command post 20 meters north of the weapon. This distance was 1/5 of the normal separation, to reduce parallax errors which would otherwise have been introduced by the scaled distances to the targets. The weapon system and observation and command posts were fixed during the experiment, with the Chaparral facing south (FM 44-3, 1977; FM 44-4, 1984).

Procedure

Each of the eight crews received two practice trials and eight test trials; one test trial for each combination of intent, weapons control status, and IFF return. Before the engagement trials began, the crews were given a simulated Operation Order specifying the mission, enemy, Primary Target Line (expected direction of attack, which was always due south), and tactical situation. They were told to search a 90 degree wide sector centered on the Primary Target Line. The crews were also told that both friendly and hostile aircraft would be attacking ground targets in their area. Specific information concerning weapons control status was passed to the crews on a trial-by-trial basis.

Prior to each trial, the experimenter selected the appropriate IFF response using the hidden Mode Select switch attached to the Chaparral IFF Subsystem Training Set. Each trial began when the crew was placed on alert and began searching for aircraft. Early warning and cueing were always given via telephone 20 to 30 seconds prior to the appearance of the aircraft. The cueing was given in the form of a clock azimuth, with the Primary Target Line designated as 12 o'clock (30 degrees to the left was 11 o'clock, 30 degrees to the right, 1 o'clock, etc.). The clock azimuth cue varied from 11 to 1 o'clock; all aircraft first appeared within 15 degrees of this azimuth. On each trial, one helicopter rose into view, hovered for 20 seconds, and then descended out of sight. The first crew member to detect the aircraft shouted "Target!" and the target's clock azimuth, and helped the squad leader to find it. The squad leader helped the senior gunner to find the aircraft, then identified the aircraft visually, aided by his binoculars. As soon as the senior gunner detected the aircraft, he centered it in the weapon sight range

ring, electronically challenged it, and listened for the IR tone which indicated that the missile's infrared seeker had acquired it. If the squad leader gave an engagement command and launching was feasible, the senior gunner "launched" the missile. After launching, the squad leader monitored the helicopter and issued another engagement order if necessary (Chaparral engagement, Parts 1 and 2). To prevent the crews from learning the identity of the helicopter on each stand, the direction each helicopter was facing was changed periodically and the type of helicopter mounted upon each stand was changed midway through the trials for each crew.

Results and Discussion

Table 1 lists the percentage of trials in which various actions occurred and their associated event times. Although the Chaparral crews detected 97% of the aircraft, the crews only destroyed 37% of the hostiles and were only able to destroy 7% of the hostiles prior to ordnance delivery. IFF return had a significant effect upon the percentage of hostiles engaged under weapons free: 100% of the hostiles returning an unknown response but only 50% of those returning a possible friend response were engaged, even though 90% of the hostiles were correctly identified visually (Chi square (1) = 5.33, $p < .05$). The fratricide rate (percentage of friends destroyed) was 16%. The RADES Chaparral crews' visual aircraft recognition performance in the field was similar to the performance of the Chaparral crews studied by TRASANA (TEA 12-81) in the classroom. Aircraft intent significantly affected the correctness of the identification decision; crews identified 90% of hostiles but only 79% of friends correctly (see Table 2).

While the overall engagement time was 17.6 seconds, gunner interrogation and firing occurred rapidly. The longest engagement events were the non-machine-aided detection and visual identification and the IR seeker lock-on events. Lock-on overlaps with visual identification. The only significant effect on any engagement event time was the three-way interaction of intent, weapons control status, and IFF return on identification time (see Table 2). For hostile helicopters, the fastest identification time (14.7 seconds) occurred in weapons free with an IFF return of unknown, and the slowest (19.5) in weapons free, possible friend. For friends, the fastest identification time (15.2 seconds) occurred in weapons free, true friend, and the slowest (21.4) in weapons tight, true friend.

TABLE 1
ENGAGEMENT PERFORMANCE

PERCENTAGE SCORES	EVENT TIMES (in seconds)
Aircraft detected.....97	Detection.....7.5
Correct handoffs.....87	IFF interrogation.....8.2
Correct IDs.....88	Track / lock-on.....16.8
Hostiles.....90	Visual ID.....17.4
Friends.....79	Missile launch.....17.6
Friends engaged 26	
Hostiles engaged.....70	
Friends killed.....16	
Hostiles killed.....37	
Hostiles releasing ordnance..93	

704

Table 2

MANOVA Results

DEPENDENT VARIABLE	INDEPENDENT VARIABLE	df	F	p	DIRECTION OF EFFECT
Detection time	Weapons Control Status (WCS)	1,56	3.40	.071	Free > Tight
ID Correctness	Intent	1,56	5.62	.021	Hostile > Friend
Decision to fire	Intent	1,56	3.50	.067	Hostile > Friend
Decision to fire	WCS x IFF	1,56	3.50	.067	Free-UNK > Free-TFPF Tight-TFPF > Free-TFPF
Time of Track	Intent x IFF	1,34	3.76	.061	Hostile-UNK > Hostile-TFPF Hostile-UNK > Friend-UNK Friend-TFPF > Friend-UNK Friend-TFPF > Hostile-TFPF
Time of ID	Intent x WCS x IFF	1,56	4.22	.045	Friend > Hostile Free > Tight, TFPF > UNK
Decision to fire	Intent x WCS x IFF	1,56	3.50	.067	Hostile > Friend Tight > Free, UNK > TFPF>

The finding that crews were more likely to correctly identify hostiles than friends means that crews are more likely to think a friendly aircraft is hostile and engage it than they are to think an enemy aircraft is a friend and fail to engage it. This may be due to the fact that a 1/5 scale Hind-D is much larger than a 1/5 scale Cobra, and therefore subtends a larger visual angle, making recognition easier. A related idea is that Hind-D's are intrinsically easier to recognize than Cobras because Hind-D's have more distinctive features. Another possible explanation has to do with response bias. The crews are predisposed to identify aircraft as hostile because when they train using the Chaparral, visual identification is not always practiced. Instead, the aircraft presented to them are assumed to be hostile. A third possible explanation for these results is that crews spend more time and/or effort learning to identify hostiles than friends.

It is possible that Weapon Control Status did not have a significant main effect because the crews usually train under Weapons Tight.

Following the Chaparral experiment, more types of helicopters, more stands, and several types of flying jet aircraft have been added to RADES. We are optimistic about the use of RADES to study a wide variety of SHORAD air defense issues in the future.

References

Baldwin, R. D. (1973). Capabilities of ground observers to locate, recognize, and estimate distances of low-flying aircraft. (HumRRO Technical Report No. 73-8). Alexandria, Virginia: The George Washington University.

Chaparral engagement of aerial targets, part 1 (CL7838F) (undated). Fort Bliss, Texas: U.S. Army Air Defense School.

Chaparral engagement of aerial targets, part 2 (CL7839F) (undated). Fort Bliss, Texas: U.S. Army Air Defense School.

Headquarters, Department of the Army (1976, March). FM 44-1, U.S. Army air defense artillery employment. Washington, D.C.

Headquarters, Department of the Army (1977, September). FM 44-3, Air defense artillery employment, Chaparral / Vulcan. Washington, D.C.

Headquarters, Department of the Army (1984, November 2). FM 44-4, Operations and training, Chaparral. Washington, D.C.

Headquarters, Department of the Army (1980, March). FM 44-6, Operations and training for Forward Area Alerting Radar (FAAR) and Target Alert Data Display Set (TADDS). Washington, D.C.

U.S. Army TRADOC Systems Analysis Activity (1981). TEA 12-81, Chaparral/ Redeye Training Subsystem Effectiveness Analysis, Volume 2, Chaparral Main Report. White Sands Missile Range, New Mexico.

Command and Control Teams: Techniques for Assessing Team Performance

Merri-Ann Cooper
Advanced Research Resources Organization

Samuel Shiflett
George Washington University

Arthur L. Korotkin
Advanced Research Resources Organization

The Air Force uses tactical command and control (C^2) systems to task and manage combat missions and responses to crisis situations. A tactical C^2 system provides: (1) central authority and coordination to determine how to use available forces, (2) a structure to send information and decisions between the commander and the forces and between adjacent units, and (3) a mechanism to originate and filter information from a number of sources to the commander. The key to the C^2 system is the effective and rapid collection, processing and transfer of information to the commander. Considerable resources, management attention and research have been spent in the topic of C^2 systems in recent years. In order to address several still unresolved issues on tactical C^2 systems, the Air Force Human Resources Laboratory began a research program focusing on personnel performance and training. This study (with monitor Lawrence Finegold) addressed one of the unresolved issues: team performance. The purposes of the project were to: identify the team characteristics of units operating C^2 systems, to study the ways currently used to evaluate C^2 operations to see if team aspects are considered, to study C^2 teams to determine if team functions could be observed, and to see if rating scales on team functions could be used for the assessment of C^2 team performance.

Method

In order to determine whether a team performance framework was appropriate for studying tactical C^2 systems, information was gathered about the basic characteristics of C^2 systems, about the team aspects of C^2 systems and about the procedures used to evaluate personnel performance in C^2 systems. Three methods were used to study tactical C^2 systems. First, the project staff reviewed Air Force documentation and research on C^2 units. Second, the project staff made field visits to observe three C^2 units: The North American Air Defense Command (NORAD)/20th Air Division at Ft. Lee, Virginia; the 728 Tactical Control and Reporting Center (CRC) at Eglin Air Force Base, Florida; and the 507th Wing Tactical Air Control Center (TACC) at Shaw Air Force Base in South Carolina. At these sites, semi-structured interviews were held with operational and training personnel. Third, the project staff observed exercises at two of the units and interviewed exercise participants and the associated exercise staff. The two exercises we observed were war simulations, Blue Flag, in which the TACC participated and a System Training Exercise for the CRC. During the exercises, the ARRO staff both identified team tasks and characteristics as well as observed and evaluated team performance using scales developed to assess team functions.

Because of the size and complexity of the C^2 units and the number of teams which make up these units, the observations and interviews focused on two small teams in each C^2 unit. Each team was selected using the following criteria: the team has a fixed structure, organization and task; the team participates in exercises; team interactions could be observed, and the teams differ in terms of being proceduralized (restricted in terms of activity options). The Weapons in the CRC represents the proceduralized type of team and the Fighter Duty Officer Team in the TACC was selected to represent a less proceduralized team.

Assessment of Team Functioning

Three approaches have been used to evaluate team performance. One is the evaluation of the quality of team's product (e.g., Bass, Farrow & Valenzi, 1979; Morgan, Coates, Kirby & Alluisi, 1984). A second is a measure of the performance of the entire system of which a team is a part (e.g., Obermayer, Vreuls, Muckler & Conway, 1974). Neither of these methods is appropriate in this study since they do not differentiate among the contribution to outcomes of individual team members, of teams, and of available resources. We decided, instead, to use a measure of team functions. The measure was based on a taxonomy initially developed by Nieva, Fleishman and Rieck (1978) from a review of the substantial literature on team performance. Draft scales measuring team functions (see Figure 1 for an example) were developed and pretested in a second study by Shiflett, Eisner, Price and Schemmer (1982) in which films of Army combat and combat support teams were observed. Naive raters who had been briefly trained about the functions, observed taped segments and used the rating scales to determine whether or not the functions were present. These raters observations were reasonably reliable.

The team taxonomy is presented in Table 1. The functions in the taxonomy are organized into five general categories: (1) Orientation involves, the generation and exchange of information concerning member abilities, the task, and the environment (resources available, opposition data, and environmental conditions); (2) Resources Distribution concerns matching task requirements to member abilities and the number of members required; (3) Timing involves pacing of activities by the team and by individual team members; (4) Response Coordination concerns the coordination of activities between team members to ensure synchronized performance; and (5) Motivation which involves energizing and directing the group (e.g., developing group norms, reinforcing performance, and resolving conflicts).

Team Characteristics of C^2 System

One of the two C^2 teams studied was the Weapons Team in the CRC. The Weapons Team has primarily responsibility for distributing aircraft to intercept unidentified aircraft and attacking hostile aircraft. The Weapons Team is headed by the weapons assignment officer (WAO) who is assisted by a weapons assignment technician. Under the WAO are subteams made up of dyads, each composed of an air weapons controller and an air weapons controller technician. Each of these dyads sits at a radar scope console and has primary responsibility for controlling aircraft in a particular area of the battle zone. The pilot is dependent primarily on that controller to get the aircraft to the target safely and efficiently. In the CRC team, most of the team functions were expressed in terms of communication activities, the

Rate the extent to which you perceived the RESPONSE COORDINATION function occurring indicating the degree to which the team coordination efforts occurred in a complex and detailed manner, requiring careful and continuous monitoring of other team member activities.

7	____	The response coordination involved <u>very complex</u> and detailed adjustment and sequencing of behavior.
6	____	
5	____	
4	____	The response coordination involved <u>moderately complex</u> adjustments and sequences of behavior.
3	____	
2	____	
1	____	The response coordination involved <u>simple</u> adjustment and sequencing of behavior.

Figure 1: Scale measuring response coordination

Table 1

Taxonomy of Team Functions

A.	Orientation Functions
	1. Information exchange regarding member resources and constraints
	2. Information exchange regarding team task and goals/mission
	3. Information exchange regarding environomental characteristics and constraints
	4. Priority assignment among tasks
B.	Resources Distribution Functions
	1. Matching member resources to task requirements
	2. Load balancing
C.	Timing Functions (Activity Pacing)
	1. General activity pacing
	2. Individually oriented activity pacing
D.	Response Coordination Functions
	1. Response sequencing
	2. Time and position coordination of responses
E.	Motivational Functions
	1. Development of team performance norms
	2. Generating acceptance of team performance norms
	3. Establishing team-level performance-rewards linkages
	4. Reinforcement of task orientation
	5. Balancing team orientation with individual competition
	6. Resolution of performance-relevant conflicts

From Shiflett, Eisner, Price, and Schemmer (1982)

720

major behavior in the CRC. The Orientation Function, defined in terms of information about intercepting a target, was observed in terms of display plotter board postings of the air situation, that are monitored by team members. A lot of time was spent on Response Distribution, as information about available resources was updated so that resources could be assigned to new targets. The Timing function was most directly seen in the attempts to ensure that the posting of the actual situation was current. Response Coordination functions occurred primarily between teams, as information flowed from the Surveillance Team, which identifies aircraft as friendly or hostile, to the Weapons Teams, to the people who update the plotter boards. Motivation functions were not in evidence, which is not surprising since they typically occur early in the groups development (Shiflett et al., 1982). One function we observed, that is not in the taxonomy, concerned the detection of errors or system monitoring.

The second team selected for study was the Fighter Duty Officer Team in the TACC. The TACC is the tactical air control element with primary responsibility for command and control of theater operations. This team has responsibility for carrying out, monitoring and adjusting tactical air operations as planned the previous day. The team receives and reviews the plan concerning the types of missions, the number of each type of mission, and the resources to be allocated. The team members plan in detail the sorties assigned to them. When the plan is changed, the team is responsible for planning new missions. It's job is to get the appropriate number of the proper type of aircraft, armed with the correct ordnance for the job, to the target and back to base or to an alternate landing site. In addition they are responsible, after the sorties are planned, to post them on the status boards in order of expected time over target. In the TACC, there was some interaction within the team, but most of the interaction observed were between team members and other teams, either in the TACC or in the cooperating units (e.g., the CRC, Army units). The senior members of the Fighters Duty Officer Team were involved in Orientation functions, by updating the team on changes in targets, intelligence and the battle situation. Resource Distribution was observed in the attempts of team members to plan sorties so that the allocated resources achieve the assigned missions. The Timing function was obviously a key aspect of the TACC, since the battle situation is time sensitive. Making sure that aircraft were allocated in a timely manner involved coordination in the team and with other teams. Response Coordination is the most critical of the team dimensions in the TACC since the TACC coordinates the activities of the offensive war. However, most of these functions are with individuals outside the primary team: with ground troops, tankers, reconnaissance, and the CRC. The motivational functions were not directly observable, although the level of motivation could be inferred from the heightened activity at certain periods. Finally, system monitoring or error detection was again observed in the attempts to obtain feedback about whether the missions proceeded as they should.

Evaluation of C^2 Performance

Several types of assessment procedures are currently used to evaluate Air Force personnel who operate C^2 systems: assessment of the performance of units during training exercises, assessments of operational readiness, tests of individual performance and knowledge, and reviews of paperwork. In the two exercises that were observed, controllers observed the exercise, identified problems in performance and provided feedback to participants.

These evaluations concerned outcomes (e.g., number of enemy aircraft destroyed) rather than behavior. Individual performance and team interactions were rarely explicitly evaluated. This may be partially the result of the few evaluators, relative to participants, present in the exercise. Since these are primarily training exercises, the focus is also on improving performance, rather than evaluation. The other evaluations (Operational Readiness Assessment, Standardization/Evaluation Report and Management Effectiveness Inspection), despite being more standardized and more intentionally evaluative, do not assess team performance or coordination.

Use of the Team Functions to Evaluate C^2 Functioning

A major purpose for studying tactical C^2 teams was to determine the adequacy of the taxonomy of team functions as a basis for evaluating C^2 performance. In general, it was found that the team functions were able to capture most team related aspects of tactical C^2 teams. All of the functions, except for motivation, defined in the taxonomy appear to be present, to some degree, in tactical C^2 teams. It became apparent that one function that occurred was not in the taxonomy. It was system monitoring, which involves the checking of all system elements for errors and ommissions.

A major problem emerging from the observation of C^2 teams involved the scales used. The scales describe the level of a function without regard for the situational requirements. Such scales may be less appropriate for evaluation than for description. In addition, in the C^2 setting, certain teams functions did not seem to differ in degree and thus were not appropriate to be rated on a seven-point scale. Rather, measures involving presence/ absence, frequency or timeliness might be more appropriate. There were two additional problems in assessing performance in tactical C^2 teams. First, the complexity of the C^2 situation made understanding and evaluation difficult for an observer who is not knowledgeable about the C^2 environment. Second, because many team functions involved communication, an observer not linked into the communication system could not effectively observe many team interactions.

References

Bass, B. M., Farrow, D., & Valenzi, E. (1979) Analyses of PROFILE data. Miami, Fl: Florida International University.

Morgan, B. B., Coates, G. D., Kirby, R. H., & Alluisi, E. A., (1984) Individual and group performance as a functions of the team training load. Human Factors, 26, 127-142.

Nieva, V. F., Fleishman, E. A., & Rieck, A. (1979) Team dimensions: Their identification, their measurement, and their relationships. Final Report, Bethesda, MD: Advanced Research Resources Organization.

Obermayer, R. W., Vreuls, D., Muckler, F. A., & Conway, E. J. (1979) Combat-ready crew performance measurement system: Final Report. Nortbridge, CA: Manned Systems Sciences, Inc.

Shiflett, S. C., Eisner, E. J., Price, S. J., & Schemmer, F. M. (1982). The definition and measurement of team functions. Final Report. Bethesda, MD: Advanced Research Resources Organization.

Item Factor Analysis of ASVAB 14

Clarence McCormick
HQ U.S. Military Entrance Processing Command

In the Fall of 1984, for a special project, we tested all the students (about 15,000) in about 50 high schools scattered around the country with the Armed Services Vocational Aptitude Battery, Form 14. This study reports the results of some item factor analyses of Form 14A. The results for Forms B and C replicated those reported here. The students were 5,260 9th-12th graders with approximately half male and half female. The analysis involved only the 8 power scales. All analyses used the SPSSX principal axes procedure with squared multiple correlation as initial estimates of the communalities, the greater than unity equivalence criterion and varimax rotations.

All Items. Phi coefficients were used rather than tetrachorics because the procedure available to us for calculating tetrachorics could not handle the 200 items. Forty-three factors were initially extracted. After rotation only five factors showed enough variables loading on them highly enough (loadings equal to or greater than .30) to define the factors. The rest were all either singlets or doublets relative to this criterion. Table 1 presents a summary of the results for these 5 factors. The table presents the number of items in each subtest which showed loadings equal to or greater than 30 on each of the 5 factors; the average and range of the difficulty (p) values for those items; the final solution eigenvalues; and percents of variance accounted for by each factor.

Table 1

Summary of Factor Analytic Results for all 200 Items

	FACTORS				
Subtest	I	II	III	IV	V
General Sciences	6			5	
Arithmetic Reasoning	4	19			
Word Knowledge	15			14	
Paragraph Comprehension	7				
Auto and Shop Information			17		
Mathematics Knowledge	2	8			3
Mechanical Comprehension	2		2		
Eletronics Information			6		
Average "p"	.75	.46	.47	.44	.23
Range of p-values	.63-.92	.30-.62	.24-.74	.22-.39	.20-.26
Eigenvalues	26.13	4.01	3.36	1.99	1.17
Percent of Variance	13.1	2.0	1.7	1.0	0.6

The first three factors are clearly the Verbal, Math, and Shop/Technical factors reported for the ASVAB when the subtest scores are factor analyzed (Ree, et al., 1982). Factor IV is another, clearly defined, Verbal factor; and Factor V is another, minimally defined, Math factor.

The two verbal factors are clearly separated on the basis of the diffculty levels of the items. The first factor consists entirely of relatively easy

items while the third factor consists of items at moderate difficulty levels. There is no overlap in the ranges of these values for the two sets. Similarly, for the two

Match factors; however, the distinction now is between moderate and difficult items. Presumably, Factor V might have been better defined if there were more relatively difficult Math items available in the pool.

Factor III, however, presents us with simply a content factor with difficulty values over the full range from easy to hard. In the case of this set of items the common variance attributable to content seems to outweigh common variance on the basis of difficulty level. Perhaps if there were more easy and hard items available in the pool for this content area we might find additional difficulty factors.

With this idea in mind we can look at the 4 Arithmetic Reasoning, the 2 Math Knowledge, and the 2 Mechanical Reasoning items which appear on Factor I. We might hypothesize, as is often done, that these items are primarily verbal in nature and thus show very close relationsips to other verbal items in the pool. Perhaps. However, it is clearly true that these items are very easy items (the p-values range from .64 to .92) like the verbal items on the factor. It seems quite reasonable to hypothesize that these items are allocated to the first factor primarily on the basis of common difficulty level variance rather than common content variance. Again, we might infer that if more easy items for these content areas had been available we might have found separate difficulty factors for them.

One inference, then, which can be made from these results is that item factor analysis can supply information useful for the construction of batteries such as the ASVAB. Given an initial pool of items with relatively large subsets balanced for difficulty level as well as for content we can construct well structured tests which should help eliminate some of the discrepancies found in the factor analytic literature where some items seem to migrate from factor to factor across different studies.

Another important implication of these results is that counselors need to bear in mind the hierarchial nature of the test scores used in the battery. At the item level we find (ignoring differences in difficulty level) three separate, orthogonal sources of variance in the content of the items. When these items are clustered together into subtests and the resultant subtest scores are factored we find the same three factors; however, now these factors are correlated. These intersubtest correlations imply that the subtest scores tend to vary similarly--either up or down--on the average. That is to say that at the level of the subtest scores we find evidence of some general reasoning factor which is not evident among the item responses. It might be useful, as Jensen (1985) points out, to examine the relationships between different levels of general ability and various training criteria. Nevertheless, the counselor should also be aware of the fact that underlying the general factor are several relatively independent sources of variation which can contribute to various differences in individual and subgroup profiles. As an example of this principal we can examine an analysis of the ASVAB composites. The example also serves to underscore the sensitivity of

factor analysis to differences in the samples of items (or scales) which are intercorrelated for the analysis.

Composite Forms. To conserve space we present here only an analysis of the Verbal Composite which is made up of the Word Knowledge, Paragraph Comprehension and General Science subtests (75 items). However, it should be noted that the same results were obtained for the other six composites used in the Student Testing Program.

Six factors were found to be interpretable after rotation. Factor I consisted of 14 WK and 4 GS easy items; and Factor II of 13 WK, 1 PC, and 1 GS moderately difficult items. Factor III consisted of 6 GS and 1 WK moderately difficult items; and Factor V of 4 GS relatively easy items. Factor IV consisted of 8 PC easy items; and Factor VI of 5 PC moderately difficult items.

Thus, when all items were factored together we obtained 3 orthogonal factors reflecting differences in common variance between the three general content areas covered by the ASVAB; and the Word Knowledge, Paragraph Comprehension, and General Science items all loaded, for the most part, on the first factor. However, when we analyse only this set of items we find that the factor breaks up into 3 orthogonal sets representing the 3 separate subtests relative to content plus a matched set representing differences in difficulty levels.

Again we might note that if the subtests had been balanced in terms of equal numbers of hard, moderate, and easy items and in terms of relatively equally high loading items within each content area the results would probably be even clearer. However, we would like to stress here the similarity between these results and those of the multiple regression studies from which these composites were constructed. Certainly, the composite scores are highly intercorrelated. Never the less, within each composite the item sums would seem to be relatively independent and can be expected to vary relatively independently for at least some individuals and for appropriately constituted subgroups. Of course, we might expect to maximize that relative independence by using the results of item factor analyses to help select an appropriately balanced set of items for each subtest.

Subtest Items. The items of each subtest were also factored to obtain evidence on the dimensionality of each of the subtests. Current opinion seems to favor the use of tetrachorics when factoring dichotomous items. It is hard to see why this should be so. The reason often given is that tetrachorics are unaffected by item difficulty levels and so will not produce so called "difficulty" factors. However, it has long been known that difficulty factors can be obtained as readily when tetrachorics are used as when phis are used (Guilford, 1941; Gourlay, 1951). It is also well known that tetrachorics are inappropriate whenever guessing is likely to be involved in the responses (Carroll, 1945; Lord, 1980) as would seem to be the case for many ASVAB items.

At any rate, we calculated item intercorrelations both as phis and as tetrachorics and compared the resulting factor solutions. Both types of correlation coefficients produced the expected "difficulty" factors. In fact, the solutions were so similar that we simply present the summary results for the analysis which used the phi coefficents (Table 2).

TABLE 2

Summary of the Factor Analytic Results for the Subtest Item

	I		II		III	
Scale	No. of Items	Average p	No. of Items	Average p	No. of Items	Average p
WK	14	.80	21	.47		
PC	8	.67	7	.43		
GS	11	.66	4	.55	10	.34
AR	5	.73	7	.59	16	.39
AS	6	.60	12	.44	6	.27
MK	10	.60	3	.43	12	.35
MC*	10	.55	4	.41	8	.36
EI	7	.56	5	.42	6	.26

*MC also exhibited a doublet with difficulty values in the .20's.

It seems clear from the table all the factors for each subtest seemed to be difficulty factors. WK and PC exhibited two factors each: one for easy, the other for moderately difficult items. The other subtests showed three factors each; one for the easier, one for the moderately difficult, and one for the more difficult items.

Examination of the content of the items on each of the factors failed to reveal any consistent trends for the factors to be related to content differences.

Thus, the major source of variance among the item responses for each of the subtests would seem to be difficulty level. Happily so; for it is just these individual differences which we wish to measure. Each of the subtests, then, can be assumed to be unidimensional varying only along a scale of difficulty. This, seems to be true even where we might not expect unidimensionality. For example GS is made up, basically, of two different items types: those which have content taken from the physical sciences and those with content from the biological sciences. Yet, these items (as we saw in the initial analysis) do not seem to be measuring achivement in these separate areas of study. Rather, they seem to be measuring a more general aptitude: an aptitude for verbal learning.

The same inference as to the basic undimensionality of the subtests can be found by looking at another index of undimensionality the ratio of the first eiginvalues to the second as compared to the ratio of the second to the third. The former ratio ranged from 2.5 to 4 (with most about 3.5) while the latter ranged from 1.2 to 2.2 (with most about 1.5).

Note: Specific statistics obtained in these factor analyses are available from the author upon request.

References

Carroll, J.B. The effect of difficulty and chance success on correlations

between items or between tests. Psychometrika, 1934, 10, 1-19.

Gourley, W. Difficulty factors arising from the use of tetrachorics correlations in factor analysis. British Journal of Psychology: Statistical Section, 1951, 4, 65-73.

Guilford, J.P. The difficulty of a test and its factor composition Psychometrika, 1941, 6, 67-77.

Jensen, A. Test Review: Armed Services Vocational Aptitude Battery. Measurement and Evaluation in Counseling and Development, 1985, 18, 32-37.

Lord, F. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum, 1980

Ree, M.J., Mullins, C.J., Mathews, J.J. & Massey, R.H. Armed Services Vocational Aptitude Battery: Item and Factor Analyses of Forms 8, 9, and 10. AFHRL-TR-81-55. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, March 1982.

717

ALTERNATE FORMS RELIABILITY OF THE ARMED SERVICES VOCATIONAL APTITUTDE BATTERY (ASVAB) FORMS 8, 9, AND 10

Brian M. Stern
Human Resources Research Organization

Leonard A. White, Hilda Wing, and Sidney A. Sachs
U.S. Army Research Institute for the Behavioral and Social Sciences

The purpose of this research is to document the reliability of examinee scores on alternate forms of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10. The ASVAB Forms 8, 9, and 10 are composed of ten subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto Shop Information (AS), Math Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). Two of these subtests, Numerical Operations and Coding Speed, are considered speeded tests. The remaining eight subtests are power tests with specific time limits for each administration. ASVAB subtests are grouped together in specific ways to form Aptitude Area composites and a measure known as the Armed Forces Qualification Test (AFQT). Scores on the AFQT are used for selection and scores on the ten Aptitude Area composites are used for occupational classification.

The ASVAB 8, 9, and 10 has six parallel forms designated as 8a, 8b, 9a, 9b, 10a, and 10b. For each of the six forms, the four subtests that comprise the AFQT consist of a unique set of items. The remaining six subtests have three parallel forms (8a, 8b), (9a, 9b), and (10a, 10b). For these six subtests, identically numbered ASVAB forms (e.g., 8a and 8b) contain the same items. Since the six forms have been shown to be equivalent (Ree, Mathews, Mullins, & Massey, 1981) the present research did not distinguish among them.

Reliability of ASVAB Subtest

Wilfong (1980) examined the test-retest reliability of the ASVAB Form 5, the DoD Student Testing Program version of the ASVAB. Test-retest reliability coefficients for six subtests ranged from .61 to .83. More recently, McCormick, Dunlap, Kennedy, and Jones (1983) investigated the stability of ASVAB 8, 9, and 10 subtests and composites using a group of 57 trainees enrolled in a Job Corps center. As part of a research project, trainees were administered parallel forms of the ASVAB on successive days. After correction for range restriction, the coefficient of stability and equivalence for ASVAB subtests ranged from .72 for PC to .88 for MK with a median reliability of .837.

The purpose of the present research was to examine the reliability of alternate forms of ASVAB 8, 9, and 10, in the Army operational environment where the test-retest interval is at least 30 days.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

1748

METHOD

Subjects

The sample consisted of 23,129 Army applicants who tested and retested on ASVAB 8, 9, and 10. For this group of applicants retesting occured in FY81. The test-retest interval ranged from 1 to 11 months.

Records of ASVAB scores are maintained by the Military Enlistment Processing Command (MEPCOM). Through coordination with MEPCOM a data file containing the test scores of applicants was made available to the Army Research Institute. The scores for each examinee were the results obtained on the two most recent ASVAB examinations. Applicants who were asked to retest for verification purposes were not included in the sample.

Of the 23,129 applicants who retested, 17,063 (74%) failed to achieve the minimum AFQT scores required for enlistment on previous testing. In the sample, 18,422 (80%) were male and 4,707 (20%) were females; 10,981 (48%) were white, 10,669 (46%) were black, and 1,479 (6%) were other than black or white.

Procedure

Pearson product moment correlations were computed between subtest scores obtained by examinees on alternate forms of the ASVAB 8, 9, and 10. However, the variance of scores on ASVAB subtests in the sample was subject to restriction in range. As a consequence of this reduction in variance, reliability estimates were reduced relative to the population values. The following formula derived by Kelley (1923) was employed to obtain reliability estimates for the unrestricted population:

$$R_{xx} = 1 - \frac{s_x^2}{S_x^2} (1 - r_{xx}),$$

where

s_x^2 = the variance in the restricted sample; r_{xx} = the reliability in the restricted sample; S_x^2 = the variance in the unrestricted population; and R_{xx} = the reliability in the unrestricted population.

The reference population was a national probability sample of men and women aged 16-23 (Office of the Assistant Secretary of Defense, 1982).

RESULTS AND DISCUSSION

Table 1 presents information pertinent to the evaluation of the reliability of ASVAB subtests. The magnitude of restriction in range can be seen by comparing SD's in the population (column 2) to those SD's obtained for the sample (column 1). Correlations between examinee scores on alternate forms of the ASVAB subtests were computed and adjusted for range restriction using the Kelley formula. The alternate form reliabilities of ASVAB subtests in the population ranged from .62 to .87 (column 4). Of the power subtests, the lowest reliabilities were obtained for the two shortest subtests, Paragraph

TABLE 1

Reliability Estimates of ASVAB Subtests

Subtest	Standard Deviation		Alternate Forms		Mean Internal Consistencies	IRT Estimates
	Sample	Population	Uncorrected	Corrected		
GS	3.74	5.01	.62	.78	.86	.86
AR	4.17	7.37	.58	.87	.91	.87
WK	5.69	7.71	.71	.84	.92	.86
PC	2.79	3.36	.47	.62	.81	.68
NO	9.10	10.99	.65	.76	.78[a]	.71
CS	13.35	16.25	.59	.72	.85[b]	.82
AI	4.75	5.55	.72	.80	.87	.83
MK	3.09	6.39	.43	.87	.87	.84
MC	4.04	5.35	.62	.78	.85	.83
EI	3.21	4.24	.53	.73	.82	.80

[a]Parallel form reliability estimate as reported in Sims & Hiatt (1981).
[b]Mean parallel form reliability estimate as reported in Wilfong (1980).

Comprehension and Electronics Information. Estimates of reliability based on the average KR-20 across forms (Ree et al., 1982) and an Item Response Theory (IRT) model (Bock & Mislevy, 1981) are also presented in Table 1. As expected, the reliability of alternate forms separated in time was somewhat less than the reliability of one form in a single testing session.

In the Army operational environment, selection and classification decisions are based on scores on ASVAB composites. Reliability increases when subtests are combined into composites due to the larger number of items in the composite measures. The alternate forms reliability of ASVAB composites in the unrestricted range exceeded .85 in all cases.

A majority of the retesters in the sample were from the lower quartile of the ability distribution. The Kelley formula used to correct reliability coefficients for range restriction is based on the assumption that error of measurement is invariant with respect to ability variations in the groups being tested. One concern here, is that there are reasons for suspecting that error variance in the group of retesters may be higher than the average error variance in the population. Errors of measurement for examinees in the lower portion of the ability distribution are often higher due to the greater prevalence of random response, carelessness, and inappropriate levels of item difficulty. Research by Bock and Mislevy (1981) indicates that ASVAB subtests provide relatively high precision for measuring aptitudes of individuals at the lower end of the ability distribution. However, error of measurement is somewhat higher at the very lowest ability levels. In addition individuals who chose to retest may be among the most highly motivated to improve their previous test scores. Thus, there are several reasons for suspecting that error variance in the sample of retesters is somewhat greater than error variance in the population. Under these conditions, the Kelley formula yields a conservative estimate of test reliability in the unrestricted population (Lord & Novick, 1968).

As a part of this research, Monte Carlo techniques were used to explore further the accuracy of the Kelley correction formula. The case considered was where the error variance in the restricted sample is greater than the error of measurement in the unrestricted population. Bivariate normal density functions were generated with known degrees of correlation. Heterogeneity of error variance was introduced by lowering the reliability of scores in the restricted sample relative to unrestricted population values. The preliminary results of this Monte Carlo work confirmed the arguments presented above. Results indicated that correction for range restriction using the Kelley formula resulted in more accurate reliability estimates. However, even the reliability coefficients adjusted for range restriction were conservative. The magnitude of conservative bias increase to the extent that error of measurment in the restricted range exceeded error of variance in the population.

To sum up, the alternate form reliability of ASVAB subtests and composites was found to be reasonably high. More work is needed to develop correction formulae that can adjust for complexities of distributions encountered in practice. Further investigation into this area is currently being conducted by the authors of this paper.

REFERENCES

Bock, R.D., & Mislevy, R.J. (1981). Profile of American Youth: Data quality analysis of the Armed Services Vocational Aptitude Battery. Chicago, IL: National Opinion Research Center.

Kelley, T.L. (1923). Statistical Methods. New York: Macmillan.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

McCormick, B.K., Dunlap, W.P., Kennedy, R.S., & Jones, M.R. (1983). The effects of practice on the Armed Services Vocational Aptitude Battery (Tech. Report 602). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Office of the Assistant Secretary of Defense. (1982). Profile of American Youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Office of the Assistant Secretary of Defense.

Ree, M.J., Mathews, J.J., Mullins, C.J., & Massey, R. (1981). Calibration of the Armed Services Vocational Aptitude Battery 8, 9, and 10 (AFHRL-TR-81-49). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Ree, R., Mullins, C., Mathews, J., & Massey, R. (1982). Armed Services Vocational Aptitude Battery: Item and factor analyses of forms 8, 9, and 10 (AFHRLTR 81-55). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Sims, W.H., & Hiatt, C.M. (1981). Validation of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7 with applications to ASVAB Forms 8, 9, and 10 (CNS 1160). Alexandria, VA: Center for Naval Analyses.

Wilfong, H.D. (1980). ASVAB: Technical supplement to the high school counselor's guide. Fort Sheridan, IL: Directorate of Testing, U.S. Military Enlistment Processing Command.

Psychometric Properties of the Safety Locus of Control Scale

John W. Jones
Chief Industrial Psychologist
The St. Paul Insurance Companies

Lisa J. Wuebker
Industrial Psychology Graduate Student
Georgia Institute of Technology

Introduction

The purpose of this study was to further validate the Safety Locus of Control Scale (Jones, 1983). The safety scale is based on the locus of control concept (Rotter, 1966), that is, that individuals either perceive themselves as being in control of their behavior and life events (internally controlled) or being controlled by others (externally controlled).

The safety scale discriminates between internal scorers and external scorers. Internal scorers believe that they are personally responsible for their safety and can take preventive steps to avoid accidents and injuries. Conversely, external scorers believe they have little or no personal control in accident prevention. External scorers believe that accidents and injuries are due to forces outside their control, such as chance, fate, or bad luck. Both the reliability and the validity of the safety scale have been documented (e.g., Jones & Foreman, 1985; Jones & Wuebker, 1984, 1985; Wuebker, Jones, & DuBois, 1985). That is, external scorers typically have more accidents and injuries than internal scorers.

In this study, safety scores were correlated with employees' self-reports of their major on-the-job accidents, along with the estimated medical costs associated with these accidents. It was specifically hypothesized that employees exhibiting low levels of safety consciousness (i.e., external scorers) would report significantly more accidents than employees exhibiting high levels of safety consciousness (i.e., internal scorers). In addition, adverse impact analyses were conducted since the safety scale has applications in employment contexts.

Method

Employees. Two hundred eighty-three (N=283) hospital employees were sampled. The sample included 58 males, 224 females, and 1 uncoded. Two hundred sixty-three employees were non-minorities, 18 were minorities (i.e., Blacks, Hispanics, Orientals), and 2 were uncoded. Finally, 165 employees were less than 40 years old, 117 employees were 40 years of age or older, and 1 employee did not report age.

This paper was presented at the 27th Annual Conference of the Military Testing Association, San Diego, California, October 21-25, 1985.

Safety Scale. All employees completed the 17-item Safety Locus of Control Scale (Jones, 1983; Jones & Wuebker, 1985). Safety scale scores can theoretically range from -17 (external scorers) to +17 (internal scorers). Five different studies support the validity of the scale (Jones, 1985). A Spearman-Brown split-half reliability was computed on the odd versus even test items (Jones & Wuebker, 1985). Obtained reliability equals .85.

Accident Criteria. Employees completed the Employee Injury Profile (Gens, Jones, & Nesbit, 1984), a checklist that assesses major on the job accidents and injuries that employees had during the past 12 months. Employees reported how many times they were injured at work, the type of injuries, and their estimate of the total medical costs for their injuries. Types of injuries assessed included: Major cuts (5 or more stiches), serious burns or scalds, fractures, dislocations, crushed parts of body, amputations, strained back and other muscles, ruptured or herniated discs, torn muscles, hernias, eye injuries, ear injuries, and concussions. Both the total number of injuries reported and the total estimated cost of the injuries served as the criterion measures.

Procedure. Employees completed both the safety scale and the injury checklist during working hours. All employees put their names on the questionnaires. Employees were informed that the study was for research purposes only and that information would not be placed in their personnel files.

Research Design and Statistics. Employees were placed in three different groups based on their safety scores. Employees scoring below the 25th percentile (N=63) were operationally defined as the "Low Safety Consciousness Group". This group had scores that would be below the acceptable cut-off score for the safety scale. Employees scoring below standards exhibited attitudes suggesting that accidents are not preventable since they are due to uncontrollable forces such as chance, fate, or bad luck. Employees scoring above the 75th percentile (N=81) were operationally defined as the "High Safety Consciousness Group". These employees exhibited attitudes suggesting a firm belief that accidents are preventable. Finally, the intermediate group of employees (N=136) formed the "Medium Safety Consciousness Group". Both accident frequency scores and estimated medical costs were analyzed as a function of level of safety consciousness. All analyses were computed with the Statistical Package for the Social Sciences-Portable Computer Version (Norusis, 1985).

Results and Discussion

Descriptive statistics are summarized in Table 1. Inspection of Table 1 shows that the average number of accidents reported per employee equals 0.71 (SD=1.43). Finer analyses show that 71.8% (N=201) of employees reported no accidents at work, 10% (N=28) reported one accident, 8% (N=23) reported two accidents, and 10% (N=28) reported anywhere from three to seven accidents during the past 12 months.

The average estimated medical costs for injuries equals $187.68 (SD=$537.35). More specifically, 71.8% (N=201) reported no medical costs, 4.6% (N=13) reported $100 in medical costs, 4.6% (N=13) reported $200 in medical costs, and 18.9% (N=53) reported anywhere from $300 to $4,300 in medical costs for the year. The accident frequency scores significantly correlated with estimated medical costs ($r[278] = .66$, $p < .001$).

Table 1

Descriptive Statistics

Variable	N	Mean	SD	Range
Safety Scale	283	1.83	6.05	-14 to 15
No. Accidents	280	0.71	1.43	0 to 7
Medical Costs	280	$187.68	537.35	0 to $4300

Reliability. A Spearman-Brown split-half (odd versus even) reliability coefficient was computed on the 283 safety scale scores. Obtained reliability equals .79. This coefficient is consistent with the Spearman-Brown coefficient of .85 that was obtained in earlier research (Jones & Wuebker, 1985).

Validity. The average number of reported accidents was higher for the Low Safety Consciousness Group compared to the Medium and High Safety Consciousness Groups. This between-groups difference was statistically significant ($F[2/277] = 4.92$, $p < .01$). Employees who have an internal safety locus of control orientation reported reliably fewer accidents than employees with an external locus of control orientation. These findings support the concurrent, criterion-related validity of the safety scale. These findings are summarized in Table 2.

Table 2

Accident Frequency as a Funtion of Safety Consciousness

Safety Consciousness	N	Mean	SD
Low	63	1.13	1.81
Medium	136	.71	1.41
High	81	.38	.99
Total	280	.71	1.43

The estimated medical costs for injuries was higher for the Low Safety Consciousness Group compared to the Medium and High Safety Consciousness Groups. This between-groups difference was statistically significant ($F[2/277] = 3.62$, $p < .03$. These results are summarized in Table 3. They also support the validty of the safety scale.

Table 3

Injury-related Medical Costs as a Function of Safety Consciousness

Safety Consciousness	N	Mean	SD
Low	63	$346.03	$805.79
Medium	136	$148.16	$387.70
High	81	$130.86	$473.79
Total	280	$187.68	$537.35

Analysis of Table 4 and Figure 1 reveals that 38% of the Low Safety Consciousness Group were involved in one or more major accidents at work, compared to 28% of the Medium Safety Consciousness Group and 21% of the High Safety Consciousness Group. This between-group difference was statistically significant using a directional Chi-squared analysis ($X2 [2] = 5.13$, $p < .04$). This finding documents that 17% more of the Low Safety Consciousness Group were involved in accidents compared to the High Safety Consciousness Group. Other factors, such as excessive lifting demands of hospital employees, obviously contributed to accidents and injuries among both internal and external scorers on the safety scale. However, the results of this study indicate that a new psychological construct of "safety locus of control" can be assessed and used to predict behavior.

Table 4

Crosstabulation: Safety Group by Accident Group

	Safety Consciousness			
	Low	Medium	High	
No Accidents	39	98	64	201
Accidents (1 or more)	24	38	17	79
	63	136	81	280

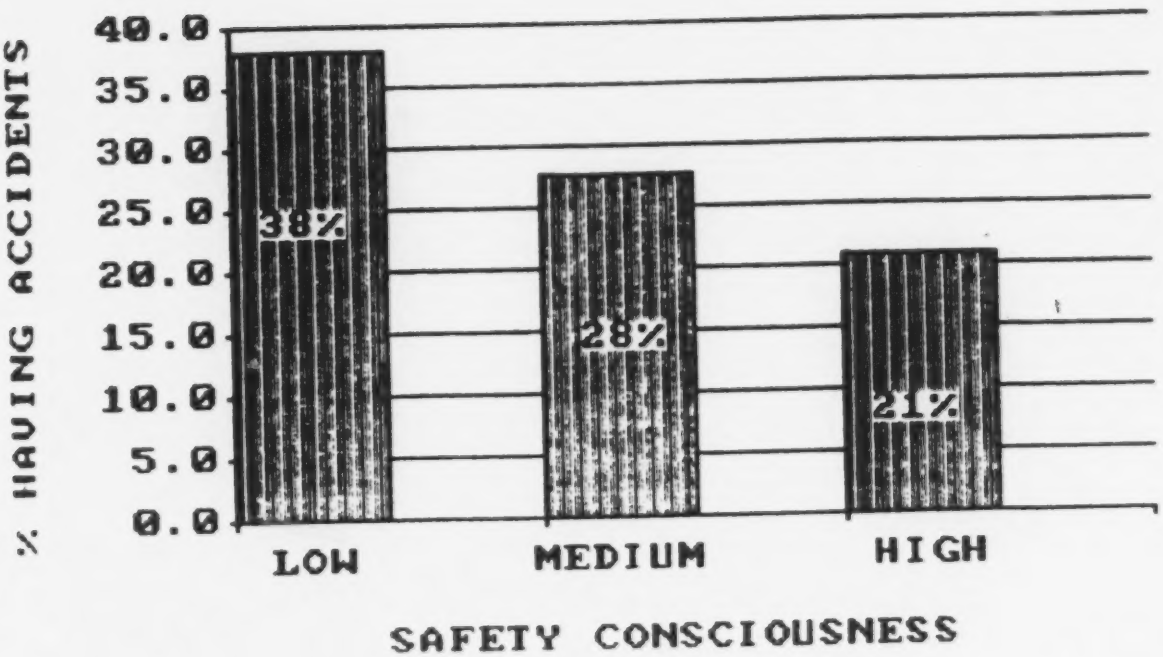


Figure 1. Percent of employees reporting one or more accidents at work as a function of safety consciousness.

Adverse Impact. Inspection of Table 5 shows that safety scores did not significantly differ as a function of sex, race, or age of employees. Analysis of Table 6 shows that the pass-fail rates for employees did not significantly differ as a function of these same demographic variables. For operational purposes, employees in the Low Safety Consciousness Group were considered to have failed the safety scale, while employess in the Medium and High Safety Consciousness Groups were considered to have passed the test.

Analysis of Table 6 shows that 20.6% of males and 22.7% of females scored below standards. In addition, 22% of non-minorities and 22.2% of minorities scored below standards. Finally, 24.8% of the employees who are less than 40 years of age scored below standards compared to 18.8% of the employees who are 40 years of age or older.

Table 5

Adverse Impact Analyses

Group		N	Safety Scale Mean	Safety Scale SD
SEX:	Males	58	3.14	6.33
	Females	224	1.15	5.96
	Total	282	1.84	6.06

F(1/280) = 3.35, p > .05, non-significant

RACE:	Non-Minorities	263	1.89	6.12
	Minorities	18	1.56	5.34
	Total	281	1.86	6.06

F(1/279) = 0.05, p >.05, non-signigicant

AGE:	<20	12	-1.25	6.22
	20-29	62	1.97	6.34
	30-39	91	1.85	6.08
	40-49	63	2.54	5.84
	50+	54	1.57	5.93
	Total	282	1.84	6.06

F(4/277) = 1.02, p > .05, non-significant

Table 6

Safety Scale Pass-Fail Rates

		Safety Consciousness Low	Medium	High	Total
SEX:	Males	12	26	20	58
	Females	51	111	62	224
	Total	63	137	82	282

Statistic: X^2 (2) = 1.03, p > .05, N.S.

RACE:	Non-minorities	58	128	77	263
	Minorities	4	9	5	18
	Total	62	137	82	281

Statistic: X^2 (2) = 0.02, p > .05, N.S.

AGE:	<20	6	3	3	12
	20-29	15	29	18	62
	30-39	20	43	28	91
	40-49	10	34	19	63
	50+	12	28	14	54
	Total	63	137	82	282

Statistic: X^2 (8) = 0.46, p >.05, N.S.

7.8

The obtained results are consistent with previous research (Jones, 1985). Employees exhibiting a high level of safety consciousness (i.e., internal scorers) are typically involved in fewer accidents and injuries at work compared to employees exhibiting a low degree of safety consciousness (i.e., external scorers). The Safety Locus of Control Scale provides industry with a standardized evaluation of safety consciousness among employees.

References

Gens, D., Jones, J.W., & Nesbit, G. (1984) The Employee Injury Profile. St. Paul, MN: The St. Paul Companies.

Jones, J.W. (1983) The Safety Locus of Control Scale. St. Paul, MN: The St. Paul Companies.

Jones, J.W. (1985) The Safety Locus of Control Scale: a review of validation research. St. Paul, MN: The St. Paul Companies.

Jones, J.W., & Foreman, R.J. (1984) Relationship of HFPSI safety scale scores to motor vehicle reports. (Technical Report) St. Paul, MN: The St. Paul Companies.

Jones, J.W., & Wuebker, L.J. (1984) The HFPSI scores of a fatally-injured construction worker. Paper presented at the Fourth Annual Construction Insurance Conference, sponsored by the International Risk Management Institute, Inc., Dallas, Texas, November 13-16.

Jones, J.W., & Wuebker, L.J. (1985) Development and validation of the Safety Locus of Control Scale. Perceptual and Motor Skills, 61, 151-161.

Norusis, M.J. (1985) SPSS/PC+: Advanced Statistics for the IBM PC/XT/AT. Chicago, IL: SPSS, Inc.

Rotter, J.B. (1966) Generalized expectancies for internal versus external locus of control. Psychological Monographs, 80 (Whole No. 609).

Wuebker, L.J., Jones, J.W., & DuBois, D. (1985) Safety locus of control and employee accidents. Proceedings of the Sixth Annual Industrial Psychology Graduate Student Conference, University of Akron, Ohio, April 12-14.

Reprints available from Dr. John W. Jones, Chief Industrial Psychologist, the St. Paul Insurance Companies, 385 Washington Street, St. Paul, MN 55102.

HEALTH LOCUS OF CONTROL BELIEFS AMONG INFANTRYMEN

LAURA W. GRIEGER
U. S. ARMY INFANTRY SCHOOL, FORT BENNING, GEORGIA

Who is responsible for one's health? The answer to this question is a matter of major concern for military leadership at every level. Optimum health and fitness are high priorities among the myriad of concerns and responsibilities of every commander. This study will explore the health locus of control beliefs among Infantrymen to discover whether their beliefs are internal or external and whether demographic factors such as religious preference, ethnic background, age, or level of education account for differences.

Background to support this study begins with a review of the related literature. Development of the Health Locus of Control (HLC) Scale, by Wallston, Wallston, Kaplan, and Maides, was as a unidimensional measure of peoples' beliefs that their health is or is not determined by their behavior (Wallston, Wallston, DeVellis, 1978). The HLC Scale was designed to yield a single score, using a Likert-type scale response, to indicate health-related behaviors as primarily one of the following two types:

a. Health externals: Individuals whose generalized expectancies that the factors which determine their health are such things as: luck, fate, chance, or powerful others, or factors over which they have little control.

b. Health internals: Individuals whose generalized beliefs that control for health is internal and that an individual stays or becomes healthy or sick as a result of his/her behavior.

From both a theoretical and empirical perspective, the exploration of health locus of control is warranted. Rotter's theory that peoples' behavior can be predicted from a knowledge of how they view a situation, their expectancies about their behavior, and how they value the outcomes that might occur as a result of their behavior in that situation (Rotter, 1966) provide the theoretical basis for this study. From the empirical perspective, an increasing number of health researchers have measured locus of control beliefs and have attempted to relate these expectancies to a host of health-related behaviors (Strickland, Wallston, and Wallston, 1978).

Levenson (1974) questioned the conceptualization of locus of control as a unidimensional construct. She reasoned that not only are internal beliefs at right angles to external beliefs, but understanding and prediction could be further improved by studying fate and chance expectations separately from external control by powerful others. Her work further resulted in the decision by K. Wallston and G. Kaplan to reconceptualize health locus of control along multidimensional lines, to develop new scales consisting only of personally worded items, and to create two equivalent forms of the health locus of control scales. Table 1 presents the items chosen for each of the three scales and form pairs for the Multidimensional Health Locus of Control (MHLC) Scales.

TABLE 1

Multidimensional Health Locus of Control (MHLC) Scales Equivalent Forms

Internal Health Locus of Control Items (IHLC)

Form A	Form B
1. If I get sick, it is my own behavior which determines how soon I get well again.	1. If I become sick, I have the power to make myself well again.

6. I am in control of my health.	6. I am directly responsible for my health.
8. When I get sick I am to blame.	8. Whatever goes wrong with my health is my own fault.
12. The main thing which affects my health is what I myself do.	12. My physical well-being depends on how well I take care of myself.
13. If I take care of myself, I can avoid illness.	13. When I feel ill, I know it is because I have not been taking care of myself properly.
17. If I take the right actions, I can stay healthy.	17. I can pretty much stay healthy by taking good care of myself.

External
Powerful Others Health Locus of Control Items
(PHLC)

Form A	Form B
3. Having regular contact with my physician is the best way for me to avoid illness.	3. If I see an excellent doctor regularly, I am less likely to have health problems.
5. Whenever I don't feel well, I should consult a medically trained professional.	5. I can only maintain my health by consulting health professionals.
7. My family has a lot to do with my becoming sick or staying healthy.	7. Other people play a big part in whether I stay healthy or become sick.
10. Health professionals control my health.	10. Health professionals keep me healthy.
14. When I recover from an illness it's usually because other people (for example, doctors, nurses, family friends) have been taking good care of me.	14. The type of care I receive from other people is what is responsible for how well I recover from an illness.
18. Regarding my health, I can only do what my doctor tells me to do.	18. Following doctor's orders to the letter is the best way for me to stay healthy.

External
Chance Health Locus of Control Items
(CHLC)

Form A	Form B
2. No matter what I do, if I am going to get sick, I will get sick.	2. Often I feel that no matter what I do, if I am going to get sick, I will get sick.
4. Most things that affect my health happen to me by accident.	4. It seems that my health is greatly influenced by accidental happenings.
9. Luck plays a big part in determining how soon I will recover from an illness.	9. When I am sick, I just have to let nature run its course.
11. My good health is largely a matter of good fortune.	11. When I stay healthy, I'm just plain lucky.
15. No matter what I do, I'm likely to get sick.	15. Even when I take care of myself, it's easy to get sick.
16. If it's meant to be, I will stay healthy.	16. When I become ill, it's a matter of fate.

Descriptive information (means, standard deviations, and alpha reliabilities) for the Multidimensional Health Locus of Control (MHLC) Scales is shown in Table 2.

TABLE 2

Descriptive Data for the Multidimensional Health Locus of Control (MHLC) Scales

Scale	# of Items	Mean	SD	Alpha
IHLC (internality)				
Form A	6	25.104	4.891	.767
Form B	6	25.304	4.646	.710
Form A & B	12	50.409	9.051	.859
PHLC (powerful others externality)				
Form A	6	19.991	5.221	.673
Form B	6	20.974	5.487	.715
Form A & B	12	40.965	10.048	.830
CHLC (chance externality)				
Form A	6	15.574	5.751	.753
Form B	6	15.461	5.204	.691
Form A & B	12	31.035	10.204	.841

Conceptualizing locus of control as a multidimensional construct, rather than a unidimensional construct, makes it difficult to think and to talk about types of individuals or situations. Wallston & Wallston (1982) propose a typology of persons based upon possible patterns of scores on the MHLC Scales. The authors readily admit the classification is highly speculative, but balance that with their intentions not to suggest that these are personality types nor do they claim that all possibilities of health locus of control beliefs are captured in the eight patterns displayed in Table 3.

TABLE 3

A Multidimensional Health Locus of Control (MHLC) Scale Typology

Type I
"Pure" Internal

	IHLC	PHLC	CHLC
High	X		
Low		X	X

Type II
"Pure" Powerful Others External

	IHLC	PHLC	CHLC
High		X	
Low	X		X

Type III
"Pure" Chance External

	IHLC	PHLC	CHLC
High			X
Low	X	X	

Type IV
Double External

	IHLC	PHLC	CHLC
High		X	X
Low	X		

Type V
Believer in Control

	IHLC	PHLC	CHLC
High	X	X	
Low			X

Type VI
Non-Existent or Rare

	IHLC	PHLC	CHLC
High	X		X
Low		X	

Type VII
"Yea-Sayer"

	IHLC	PHLC	CHLC
High	X	X	X
Low			

Type VIII
"Nay-Sayer"

	IHLC	PHLC	CHLC
High			
Low	X	X	X

The usefulness in the typology may be in recognition of the individual differences among health beliefs. Perhaps a revolutionary change needs to take place in the socialization of commanders and health professionals to

the realization that individuals may have to be dealt with differently in preventive health behaviors, adherence to medical regimens, response to symptoms, and interaction with health care settings.

There are no hypothesis statements in this research. The purpose of this research is to gather demographic data in a target population heretofore unstudied and unreported in MHLC Scale literature. The findings will serve as a basis for future theoretical and empirical research: use of the Multidimensional Health Locus of Control (MHLC) Scale as a dependent variable relating health locus of control to demographic variables among Infantrymen.

Some of the considerations leading to the conduct of this research are:

1. What are the demographic characteristics of this target population?
2. What are the results of the MHLC Scales within this population?
3. Do the demographic variables form any recognizable patterns with the MHLC Scales?
4. Do the results of the typology form any recognizable patterns with the demographic variables?
5. What hypotheses can be formulated for future research, based on the differences and/or similarities of patterns?
6. Can recommendations, based on the results of this research, be made to commanders and managers of human resources within the Infantry to promote health-related behaviors?

METHOD

Subjects:

The target population for this study consists of the annual student load in the subject area of leadership at the U. S. Army Infantry School which includes the courses and number of students displayed at Table 4.

TABLE 4

U.S. Army Infantry School Courses and Projected Student Load

Course Title	Projected Annual Student Load
Officer Candidate School	1,200
Infantry Officer Basic	2,200
Infantry Officer Advanced	1,025
Maneuver Combat Arms Advanced Noncommissioned Officer	1,230
Total	5,655

The unrefined sample consists of one iteration of each course, 14.32 percent of the projected total annual student load. The subjects are not identified by name nor by class roster number in order to provide complete anonymity. The sample is further refined by identifying foreign students and non-Infantry and eliminating their questionnaires from the data.

The refined sample consists of 577 students, 10.20 percent of the projected total annual student load and 71.23 percent of the total number of questionnaires distributed.

The high rate of useable questionnaires may be attributed to the fact that the questionnaire was administered by instructors during class and although completion was voluntary, the classroom setting may have provided impetus to complete the questionnaire.

The questionnaires discarded were rejected for the reasons shown in Table 5.

TABLE 5

Reason for Questionnaire Rejection

Demographic Portion	MHLC Scale Position
Incomplete	Incomplete
Selection of more than one response per item	Handwritten responses
Handwritten responses	Selection of more than one response per item
Foreign students	
Women (Officer Candidate course only)	

Materials:

Forms A and B of the MHLC Scale are used to gather health locus of control data. Demographic data items displayed at Table 5 are compiled from previously validated items used by the Directorate of Evaluation and Standardization for use with the same target population.

TABLE 6

Demographic Variables

Age	Ethnic background
Pay grade	Educational level
Religious preference	Course attending

Procedure and Design:

Data collection is conducted in the following manner. During classroom leadership instruction, instructors explain the purpose of the questionnaire as part of research data collection and request the students not provide their name nor roster number to insure complete anonymity. The students are instructed to read and follow the written instructions provided on the questionnaire. Completed questionnaires are collected by the instructors and returned to the researcher for tabulation.

The questionnaires are screened. Foreign students, non-Infantrymen, women candidates in Officer Candidate School, mismarked, and incomplete questionnaires are removed.

Scoring consists of three phases. First, each MHLC Scale is scored. Secondly, median splits are determined for each of the three MHLC Scales. Thirdly, each subject is type classified into one of the 8 "types" according to their pattern of being above ("high") or below ("low") the median splits on the scales. Any score falling on the median is decided as above or below by a coin toss, heads=above and tails=below.

Each subject's scored data and demographic variables are now ready to be recorded on ADP transcript paper for key punch, verification, and duplication of the data deck.

Results and Conclusions:

The results are a compilation of frequency counts. The demographic variables results are:

Age (in years):

Under 20		20-24		25-29		30-34		35 & Over		Sample Size
0	+	172	+	253	+	131	+	21	=	577

Pay Grade:

E5		E6		E7		O1		O2		O3		Sample Size
79	+	87	+	0	+	195	+	172	+	44	=	577

Religious Preference:

Protestant		Roman Catholic		Jewish		Other		Sample Size
197		+ 223		+ 0		+ 157	=	577

Ethnic Background:

Black		White		Hispanic		Asian Amer/ Pac Is		Amer Indian/ Alaskan Nat		Other		Sample Size
167	+	390	+	16		+ 2		+ 0		+ 2	=	577

Highest Civilian Education Degree:

H.S.Dip/Equiv	A.S.	B.A./B.S.	Master	Ph.D	Other	Sample Size
117	+ 20 +	359	+ 54 +	0	+ 27 =	577

Course Attending:

Off. Cand	Off. Basic	Off. Adv.	NCO Adv.	Sample Size
81	+ 195 +	216	+ 85 =	577

Results for the MHLC Scales are displayed next. A subject was classifield into a category by their highest score among the three scales.

Internal	External: Chance	External: Powerful Other	Sample Size
521	+ 44 +	12	= 577

If a subject scored the same on two or more scales the decision was made to discard that subject.

The results of the typology are as follows:

1	2	3	4	5	6	7	8	Sample Size
207	+ 21 +	53	+ 87 +	45	+ 73 +	39	+ 52 =	577

The frequency counts reveal some interesting findings. The next step will be to apply the statistical technique of discriminant function analysis to the data using the Statistical Package for Social Science^{X} (SPSS), 1983 to determine if a prediction equation results. Results of the research are programmed for 2d Qtr FY86. You are invited to contact the author if interested.

Disclaimer: The content of this paper is the sole responsibility and opinion of the author.

REFERENCES

Levenson, H. (1974). Activism and powerful others: Distinctions within the concept of internal - external control. Journal of Personality Assessment 38; 377-383.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80. (Whole No. 609).

Strickland, B.R. (1978). Internal - external expectancies and health-related behaviors. Journal of Consulting and Clinical Psychology, 46, 1192-1211.

Wallston, K.A., Wallston, B.S., & DeVellis, R. (1978). Development of the Multidimensional Health Locus of Control (MHLC) Scales. Health Education Monographs, 6, 160-170.

Wallston, K.A., Wallston, B.S. (1981). Health Locus of Control Scales. In H. Lefcourt (Ed.), Research with the locus of control construct. NY: Academic Press, 1, 189-243.

Wallston, K.A., & Wallston, B.S. (1982). Who is responsible for your health? The construct of health locus of control. In G. Sanders & J. Suls (Eds.), Social Psychology of Health and Illness. Hillsdale, NJ: Erlbaum, 65-95.

Wallston, K.A., Smith, R.A., King, J.E., Forsberg, P.R., Wallston, B.S., & Nagy, V.T. (1983). Expectancies about control over health: Relationship to desire for control of health care. Personality and Social Psychology Bulletin, 9, 377-385.

Wallston, B.S., & Wallston, K.A. (1984). Social psychological models of health behavior: An examination and integration. Chapter in A. Baum, S. Taylor, & J. E. Singer (Eds.), Handbook of Psychology and Health, Volume 4: Social Aspects of Health. Hillsdale, NJ: Erlbaum.

Classifying Military Offenders: Application of the Megargee MMPI Typology

Mark L. Paris and Gary E. Brown
United States Disciplinary Barracks

Megargee and Bohn (1979) developed an inmate classification system utilizing the Minnesota Multiphasic Personality Inventory (MMPI). The system was originally designed to enable clinicians to assign individual MMPI profiles to prototypical classifications, to facilitate empirical establishment of relationships between classifications and probable future behaviors, and to enhance treatment of young offenders by targeting specific treatment modalities to appropriate inmate profile types based on the system.

Megargee and Bohn developed their taxonomy in three phases: (1) a statistical analysis of three samples of 100 MMPI protocols to determine the number of naturally occurring subgroups; (2) the matching of subgroups from each sample with each other, and the writing of rules to describe the larger groups resulting from the matching; and (3) the classification of a new sample of protocols according to the new rules.

Eventually, a large, broad-based study of the new system was undertaken as part of an extensive longitudinal research program at the Federal Correctional Institution, Tallahassee, Florida (Megargee, 1974). Of 1214 men subjected to classification by the system, 770 (63.4%) were classified through the use of computer programs which employed profile-type rules. The remaining 394 (32.4%) were made up of ties among classifications (248), possibly invalid profiles (79) or profiles which did not meet the criteria for inclusion in any of the classifications (117). Ultimately, all but 50 cases were classified, and these remaining 1164 individuals were studied along such dimensions as academic and intellectual level, characteristics of the developmental family, measures of interpersonal relations and adjustment, psychologists' objective measures, and various measures of institutional adjustment, including work performance, number of disciplinary infractions, time spent in Maximum custody and detail evaluations. In addition, recidivism data were also studied.

Significant relationships were established between MMPI classifications and various measures of adjustment, recidivism and personality. Furthermore, Megargee and Bohn (1979) proceeded to make treatment recommendations specific to the various classifications, based on their understanding of the implications of the data which had been gathered for each classification.

In an attempt to extend Megargee and Bohn's work to the military correctional setting, an MMPI inmate classification system was begun in February 1985 at the United States Disciplinary Barracks, Fort Leavenworth, Kansas.

Method

The computer program used to classify the inmates at the United States Disciplinary Barracks (USDB) was the one written by Meyer and Megargee (1977). Inmates were administered the MMPI (all 566 items) on Monday of the second week of their incarceration at the institution. They were informed that testing, as part of their treatment, was required, but that they would receive test feedback and that the results would be seen and used only by mental health personnel, usually their own counselor.

The MMPI's were computer-scored, with the resultant Megargee classifications generated. The program either printed the raw scores, T scores and Welsh code for each individual, as well as the appropriate classification (Able, Baker, Charlie, Delta, Easy, Foxtrot, George, How, Item, or Jupiter); or else, it would print out the profile. The profile would be printed out in the event that (1) the profile was tied among several classifications, (2) it fulfilled the criteria for none of the classifications, or (3) F>99T. In these cases, the program left it open to the clinician's discretion to break ties, and otherwise classify questionable profiles.

Interpretive reports were created for each classification by utilizing the empirical data of Megargee and Bohn (1979). These reports were made available to the individual's future counselor for his/her use in treatment planning.

Results

Data were crosstabulated, specific offense by Megargee classification, and comparisons were made between frequencies of classification found by Megargee (Megargee, 1974) and by ourselves. By chi-square analysis, three comparisons were found to be statistically significant. In general, however, although our sample size was 1/3 that of Megargee's, Table 1 shows that the pattern of similarity was quite strong.

Table 1

USDB vs FCI inmate classification (percentages)

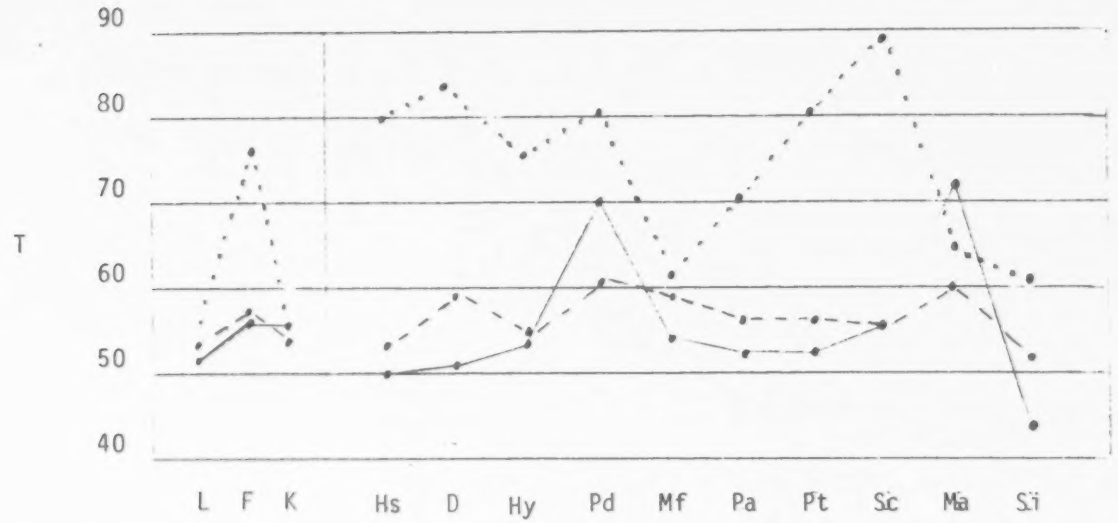
Classification	USDB (N=390)	FCI (N=1214)
Able	12.3	13.3
Baker	1.5	1.8
Charlie	4.9	4.6
Delta	5.1	7.0
Easy	3.8	3.2
Foxtrot**	1.5	5.7
George	5.6	5.9
How	5.9	7.7
Item**	25.6	13.0
Jupiter	1.3	1.1
F>99T*	2.6	6.5
No criteria met	7.4	9.6
Ties	22.3	20.4

*p<.001
**P<.0001

Inspection of Table 1 reveals that Item, the most psychopathologically benign profile in the typology, is twice as likely to be found in the USDB population as in that of the civilian FCI. Additionally, more than twice as many possibly invalid profiles are found in the FCI group as in that of the USDB. Finally, Foxtrot, the third most psychopathological of the profile types is found more than three times more frequently in the FCI as in the USDB populations.

In addition, comparison of the number of subjects falling into the various classifications revealed a significant effect among the cells, $X^2 = 301.89$, df 9, $p < .0001$. It is noteworthy that the modal profile type, Item, is essentially flat, with the mean T-scores falling in sub-clinical ranges; the next most frequent profile in the sample, Able, is the "4-9" profile type that is traditionally associated with correctional populations (see Fig. 1).

Figure 1
Mean profiles for types Able, How and Item
(after Megargee and Bohn, 1979)



Item ------
Able ———
How . . . .

When classifications were examined with respect to specific offenses (see Table 2), it was found that among three of six offense types studied, Item was found significantly more frequently than the next most frequent type (Item was also found to be the most frequent type for all other offenses, though not to a statistically significant degree). With respect to drug offenses, a bimodal distribution was found: while the frequency of Item was significantly greater than the next most frequent type, Able ($X^2 = 3.93$, $p = .036$), the frequency of Able was found to be significantly greater than the next most frequent type, Charlie ($X^2 = 6.59$, $p = .004$). Similarly, re: rape, a significant difference was found between the frequency of Item and those of either Delta or How ($X^2 = 8.45$, $p = .001$); however, Item did not differ significantly from Able, and Able did not differ from Delta or How. Thus, taken together, Item and Able tend to predominate among the other groups. Among child sex offenders, Item again significantly differed from its closest competitor, How ($X^2 = 5.84$, $p = .008$); How did not differ significantly from the next most frequent group, Delta; however, the data suggest that such a bimodal effect might be revealed as the sample size increases.

Table 2
Breakdown of selected types by offense (percentage)

	Item	Easy	Able	Delta	Charlie	How
Murder (26)	23.1	7.7	15.4	7.7	3.8	3.8
Assault (37)	16.2	5.4	2.7	5.4	2.7	13.5
Rape (54)	25.9	3.7	14.8	5.6	3.7	5.6
Child (58)	32.8	1.7	3.4	8.6	3.4	13.8
Larceny (52)	21.2	5.8	13.5	5.8	3.8	1.9
Drugs (126)	27.0	4.0	16.7	2.4	6.3	2.4

Discussion

Megargee, et al. set out, with their MMPI classification system, to create a method of predicting behavior within correctional populations. Inherent in this project was the implication that the phenomenon of "classifiability" of prison inmates should be a generalizable one; such an hypothesis could be strengthened by fairly consistent findings of profile type frequency distributions among various correctional institutions. The present pattern of data, with a few exceptions, replicates the findings of Megargee and Bohn (1979) with respect to the relative levels of psychopathology in our inmate population. Collectively, the data indicate that in military as in civilian correctional settings, a huge majority of inmates' personality profiles fall into one of ten basic types, but that in a military institution, one is less likely to find measurable psychopathology or attempts at dissimulation. It is the "flat" or basically normal profile that is the modal finding in military confinement, followed by the expected characterological "4-9" profile of Able. Although this pattern of findings was fairly consistent over several offenses, the child sex offender continues to challenge our expectations; such offenders generally present Item-type profiles, but are increasingly yielding high levels of psychopathology as well, suggesting a complex branching of etiologies going in different directions.

Future research issues will consider therapeutic implications of labeling/classification; that is, whether the process of labeling for therapy is facilitative of a self-fulfilling prophecy would seem of paramount importance if a classification system such as this is to be used for anything other than research, placement or security purposes.

References

Megargee, E.I. (1974). Applied psychological research in a correctional setting. Criminal Justice and Behavior, 4, 211-216.

Megargee, E.I. & Bohn, M.J. (1979). Classifying Criminal Offenders. Beverly Hills, California: Sage Publications, Inc.

Meyer, J. Jr., & Megargee, E.I. (1977). A new classification system for criminal offenders, II: Initial development of the system. Criminal Justice and Behavior, 4, 115-124.

The views expressed in this article do not necessarily represent those of the USDB, the U.S. Army or the Department of Defense.

AUTOMATED SCHEDULING OF ARMY UNIT TRAINING

Dwight J. Goehring
Roland J. Hart

U.S. Army Research Institute Field Unit
Presidio of Monterey, California

Computer assistance can reduce the complexity of military training scheduling for planners and lead to better planning of training. Improved training planning can make more efficient use of time and resources, and thus contribute to improving readiness. However, in order to be helpful, computer automated assistance of training scheduling must accommodate the full diversity of the military problem. Our objective is to explore the feasibility of a particular heuristic approach called simulated annealing, for solving the Army scheduling problem.

I. The Army Training Schedule Problem

The Army training scheduling problem is complex because it occurs across different echelons and involves three different types of calendars/schedules. Priorities of higher echelons almost always predominate over those of lower echelons. In addition, many constraints exist for the Army training scheduling problem. Training and non-training tasks require time and resources. Resources can be either reusable (e.g., ranges) or expendable (e.g., ammunition). Further, activities may have temporal constraints, such as requiring some tasks to be trained ahead of others, or co-occurrence constraints requiring some tasks to be trained simultaneously with others. There are also command priority constraints at each echelon which affect scheduling requirements. In addition, the priority given training varies at different times. Automated scheduling must accommodate these hetereogenous constraints.

II. Use of Simulated Annealing to Solve the Scheduling Problem

An optimization methodology is needed to solve the Army training schedule problem. An optimization methodology provides (a) a definition of what constitutes a good schedule, (b) a way of measuring numerically how good a schedule is, and (c) a method to search for numerically good schedules. Simulated annealing (Kirkpatrick, Gelatt & Vecchi, 1983) was selected as an appropriate method because of the match between the capabilities of the methodology and the characteristics of the scheduling problem. First, it can handle large problems. The training scheduling problem for an entire Army division is large. In addition, it can handle virtually any type of constraint. This feature is important because the Army training scheduling problem includes a wide variety of constraints. The approach is flexible in the sense that constraints can be changed or added and the method can still work. Flexibility is important because frequent modification is expected in an operational environment. Another feature of simulated annealing that matches the Army training scheduling problem is size of the solution space. Application of simulated annealing is appropriate when multiple "good" solutions are acceptable as opposed to a uniquely "best" solution.

Further, for large combinational problems like the scheduling problem, it becomes impossible to check all possible combinations, in search of the one combination that minimizes the cost function, because of time requirements. The time requirements for such problems increase exponentially with N (i.e., number of activities to be scheduled). Rather than search all combinations, optimization methods use heuristics to minimize cost functions. Such heuristics can be generally classified as either "divide-and-conquer" or iterative improvement strategies (Kirkpatrick, Gelatt & Becchi, 1983). Simulated annealing combines elements of both strategies. As applied to scheduling, the problem is divided into subproblems (involving common levels of "importance") and iterative improvement occurs through the application of the annealing concept.

Finally, the simulated annealing algorithm does not require a set computer time to obtain a solution. The solution will get better on the average the longer the program runs. A computer run can be terminated when a point of diminishing returns is achieved.

The simulated annealing heuristic operates by analogy to annealing in physical systems. Annealing in a physical system involves repeated heating and cooling of liquids or solids. When a liquid is hot there is a rapid random movement of the molecules making up the liquid. Heating a liquid increases the rapidity of the random movement of molecules. By contrast, cooling a liquid slows the random movement of molecules until they reach the point where they are frozen in place. Molecules frozen in place are considered "cold."

As applied to a scheduling system, the molecules are activities or events that must be scheduled. Random movement involves the random assignment of activities to times. The concept of temperature is linked to the cost function, and refers to the degree of random movement permitted relative to the cost function. High temperature is associated with high values of the cost function, low temperature is associated with low values of the cost function. High temperature means that random assignment of activities to times is permitted no matter how large the cost function gets as a result of the assignments. That is, assignments are not constrained by the cost function. On the other hand, cold temperature means that only some random assignments of activities to times are permitted, namely, assignments that reduce or at least do not increase the cost function.

The operation of the simulated annealing heuristic involves successively "heating" and "cooling" the system in search of a "good" schedule defined by a low cost function. The process of successively "heating" and "cooling" the system is accomplished to avoid the problem of getting stuck in local minimums. This problem is a common one for problems based on iterative improvement (e.g., greedy algorithms). The simulated annealing heuristic requires the creation of an appropriate "annealing schedule" involving the extent and frequency of heating and cooling.

In addition to creating an appropriate annealing schedule, an algorithm is necessary to cool the system. (Creating a hot system is not difficult since heating simply entails unconstrained random assignment of activities to times.). Cooling the system, however, requires development of an algorithm, created especially for the problem at hand, that constrains the random

assignments permitted to those producing successively lower values of the cost function, as the system is cooled. In order to "cool" a scheduling system, those tasks that have the greatest potential for increasing the cost function are scheduled first before the schedule gets too full to include them. Then tasks that have less potential for increasing the cost function are successively scheduled. Unconstrained tasks that minimally add to the cost function are scheduled last. This strategy increases the probability that important activities and constraints are accommodated in the resultant schedules.

III. Illustrative Application of Simulated Annealing

The Army scheduling problem was delimited by focusing on scheduling at the battalion level. This level was selected because it contains scheduling features from all echelons. If the annealing heuristic is effective at this level, one can assume that it can be applied at all levels with the appropriate modifications for each echelon.

All five companies within a single battalion are scheduled simultaneously in the test program. These five companies inherit training activities and events from the Long- and Short-Range Calendar created at higher echelons. The test program uses one-hour units of time for activities and forty-hour weeks. The initial test program has only one annealing schedule for "cooling" the system. This schedule flows from the top to bottom following the activity importance criterion.

The test program employs all of the important categories of scheduling constraints: (a) inheritance of activities from higher echelons and "fixing" activity times, (b) interschedule conflicts of both renewable and expendable resources, (c) company-level training activity priorities, (d) different unit priorities assigned to different companies, and (e) intra-schedule conflicts based on temporal constraints for activities (i.e., before/after, immediately before/immediately after), and temporal ordering of sequences of contiguous activities.

The simulated annealing test program was written in FORTRAN 77 on a VAX 11/780 computer. An overview of the data structures and flow of control of the simulated annealing test program is shown in Figure 1. Blocks A through E in Figure 1 depict five data structures containing the necessary input information.

Initial company-level Training Schedules developed by company commanders are simulated in Blocks F, G, and H of Figure 1. These schedules identify tasks for the week, assign priorities to the tasks ("high" versus "regular" priority), and place activities in temporal order consistent with prerequisite relationships (before/after). These initial schedules represent recommendations from a lower echelon for training schedules that are passed upward. At battalion level, resourcing is accomplished and conflicts are resolved. As depicted in Figure 1, this conflict resolution is accomplished by the passage of the initially recommended schedules to the simulated annealing algorithm in Block I. The simulated annealing algorithm produces the final set of training schedules for the companies within a battalion. In addition, lists of

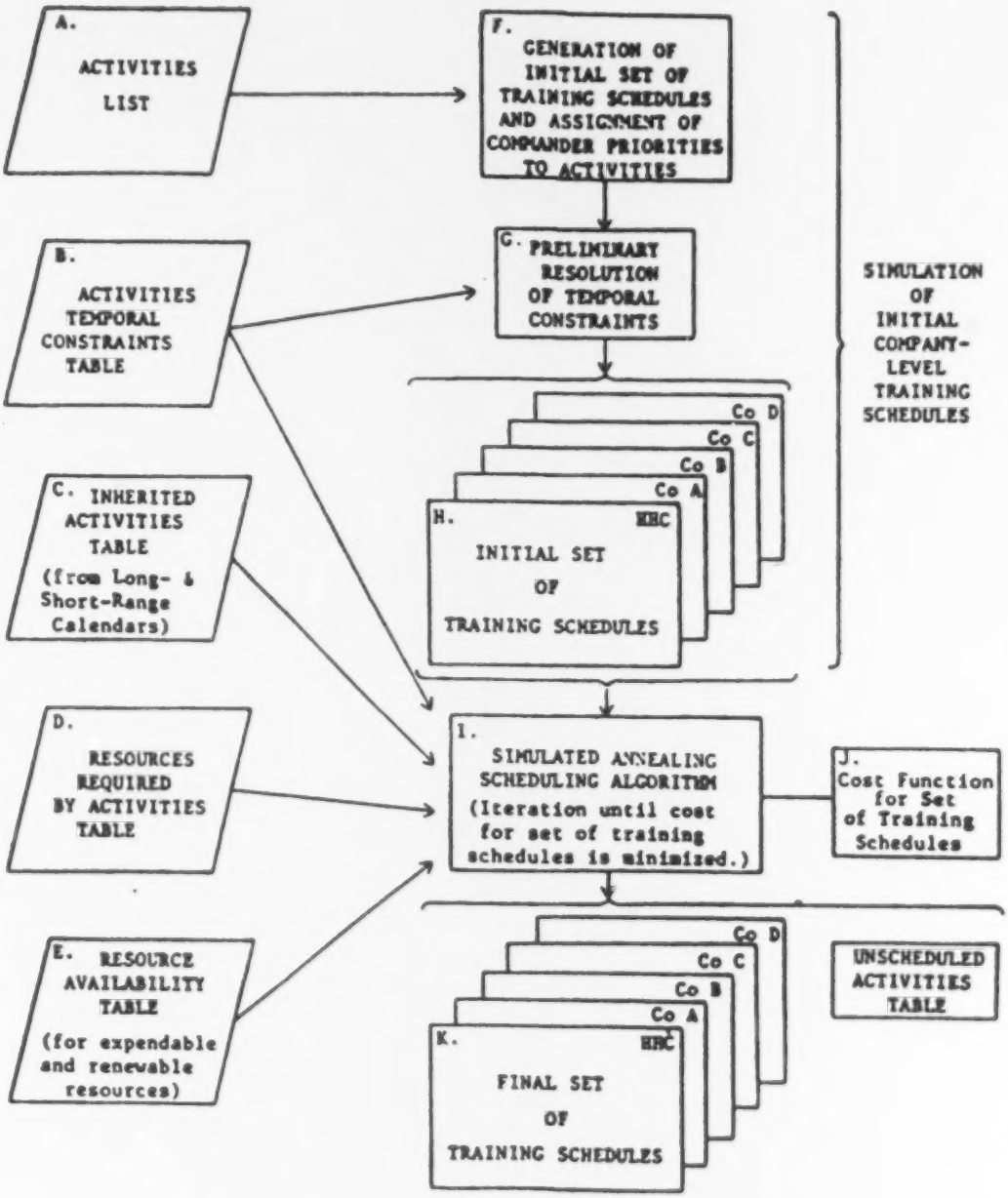


Figure 1. Data Structures and Flow of Control for Simulated Annealing Test Program

714

activities from the initially recommended schedules that could not be accommodated in the final schedules are given along with their level of importance as defined by their contribution to the cost function.

The cost function for the test program was formulated on the assumption that it is costly to fail to schedule activities. Further, it is more costly to fail to schedule longer than shorter activities and more important than less important activities.

More Specifically, let C represent cost and K represent weights reflecting the relative importance of unscheduled activities. Let U represent the time associated with unscheduled activities. Let i represent the activity number among m unscheduled activities within a Training Schedule, and let j represent the Training Schedule number that varies from 1 to n, the number of Training Schedules. Since there is a Training Schedule for each company, there are five Training Schedules per battalion ($n = 5$). Given these definitions, the cost function can be written by Equation 1 as:

$$C = \sum_{j=1}^{n} \sum_{i=1}^{m} K_{ij} U_{ij}$$

In order to create the algorithm that "cools" the system, activities need to be partitioned and sequenced for scheduling. In the test program, activities were partitioned and sequenced by the echelon level of the scheduling decision. Since activities appearing on the Long- and Short-Range Calendars are created above the battalion level, they are inherited at the battalion level and are scheduled first. (See Block 2 of Figure 2.). For the test program, it is assumed that resource conflicts and temporal constraints of inherited activities have been previously resolved. Inter-schedule conflicts between company-level Training Schedules are resolved next as shown in Block 3 of Figure 2. These conflicts generally involve conflicts over resources. Company-level training priorities (assigned at company level) are scheduled next as shown in Block 4. Then, temporal constraints yielding intra-schedule conflicts are resolved. Finally, any free blocks of time are filled with unconstrained activities yielding an end set of Training Schedules. (This decision-making sequence is an oversimplification of the real scheduling process in some respects, excluding some details, but it is sufficiently accurate and complete for describing the test program).

The heating and cooling process occurs iteratively for all company-level Training Schedules in a battalion, in search of a battalion set of Training Schedules with a low overall cost function. The weights assigned to the cost function (K_{ij}), that determine the importance of unscheduled activities, correspond in order to the "cooling" sequence just outlined.

The operation of the algorithm was sufficiently successful in simulating the performance of Army personnel experienced in training to suggest that the approach is a promising one warranting further research.

IV. Reference

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220, 671-680.

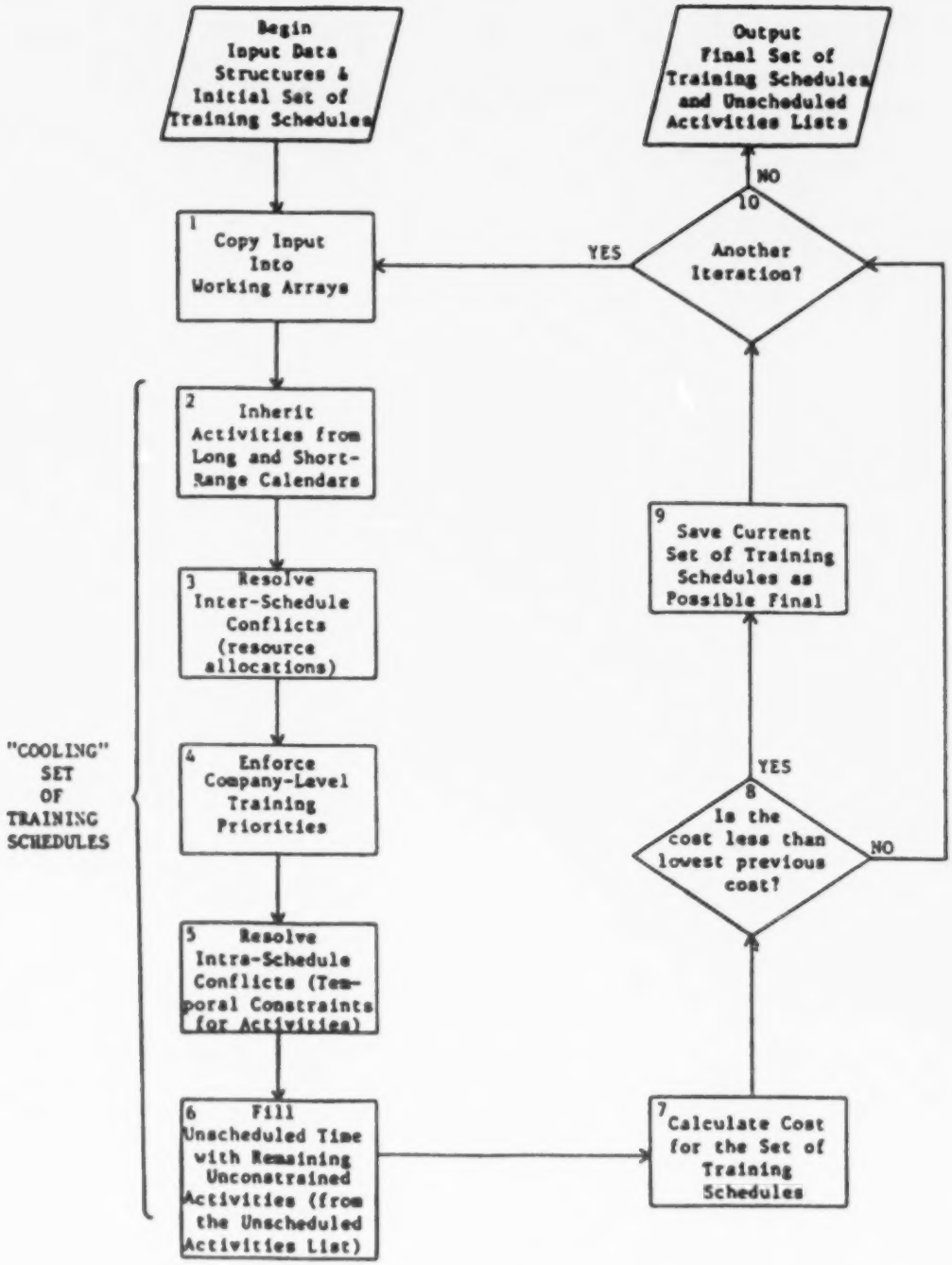


Figure 2. Operation of the Simulated Annealing Scheduling Algorithm

Supporting the Changing Role of Army Collective Training Developers

Larry L. Meliza
U.S. Army Research Institute Field Unit
Presidio of Monterey, California

Army branch schools prepare Army Training and Evaluation Program (ARTEP) documents as guides to unit collective training. The design of ARTEP documents has evolved considerably in their short, ten-year history to more effectively meet the information needs of ARTEP users. The goal of this paper is to describe 1) the evolution of the ARTEP and 2) efforts to address the growing complexity of the ARTEP developer's job through the careful application of computer technology.

Evolution of ARTEP Documents

ARTEP documents were initially designed to provide "what to train" guidance by outlining potential training requirements. These documents identify the missions and subordinate collective tasks (i.e., tasks requiring two or more individuals to perform) which a unit (e.g., mechanized infantry battalion) and its subordinate elements (e.g., companies, platoons, squads and teams) might be expected to perform. Each collective task is "defined" by indicating the tactical situation calling for its performance (conditions) and by providing task performance guidance (referred to as Standards). ARTEP users are responsible for selecting the "what to train" guidance which applies to their particular unit and combining this guidance with separately provided "how to train" guidance to develop unit training plans. During three evolutionaly phases, the ARTEP design concept has been expanded to incorporate the application of "how to train" guidance to specific missions and tasks, while leaving ARTEP users with the flexibility to select training requirements appropriate to their particular unit. (TRADOC PAM 310-8, 1981; TRADOC REG 310-2, 1982; TRADOC REG 310-2 (Test-Revised, 1985).

In the first phase of ARTEP evolution, collective/individual task matrices were added to the ARTEP document. These matrices identify individual skills prerequisites to training on specific collective tasks at the smallest unit levels (e.g., squad and platoon). This addition to the ARTEP provided the framework of a progressive unit training plan by beginning to integrate individual skills training and collective training.

The second phase of ARTEP evolution involved adding "Drill" documents to further support a progressive/building block approach to unit training at the smallest unit levels (e.g., squad and platoon) (Hiller, Hardy and Meliza, 1984). Drills are training exercises which address tasks/subtasks warranting repetitive practice to the point of overlearning. Training on drills is intended to prepare a unit for training on the more complex missions/tasks in which the drills are embedded. Since one of the criteria for selecting tasks/subtasks to be addressed by Drills is that they have broad applicability across missions, Drills have high progressive training value. Drill documents provide specific "how to train" guidance for each Drill including "set-up directions", resource requirements and "coaching points". In addition, Drill

documents provide Drill/Mission Task applicability matrices and individual skills/Drill matrices to assist ARTEP users with integrating Drills into unit training plans.

The third, on-going phase of ARTEP evolution involves the addition of ARTEP Mission Training Plan (AMTP) documents. AMTP documents represent a further extension of the progressive training plan concept by defining building block exercises called "Situation Training Exercises (STXs)" for platoon-level and above. STXs address training requirements within portions of missions (i.e., STXs often address two or more collective tasks). Many STXs are applicable to more than one mission and thus these exercises tend to have high progressive training value.

As in the case of "drills", AMTPs provide specific "how to train" guidance for each STX and matrices/figures illustrating descriptive unit training plans. In addition, STXs also provide guidance for training prerequisite leader skills.

To gain a more complete picture of how AMTP documents provide a progressive unit training plan it is important to consider that a particular unit type is intended to have different AMTP documents for various echelons. For example, a mechanized infantry battalion would have separate AMTP documents for battalion, company and platoon level. The platoon-level AMTP is intended to help a unit progress from individual skills training of soldiers to company level training, while the company level AMTP document is intended to help a unit progress from platoon level training to battalion level training.

Incorporating applications of "how to train" guidance into the ARTEP removes certain analytical tasks from the shoulder of the ARTEP user, and it helps to standardize collective training across units. Without the incorporation of such guidance, it had been necessary for each unit trainer to independently develop their own Drills, STXs and progressive unit training plans. This was a large burden for unit personnel and the quality of training programs was therefore uneven.

The Changing Role of the ARTEP Developer

The current phase of ARTEP evolution is at an experimental stage. Each of eighteen schools has prepared a prototype platoon-level AMTP to support Army-wide field testing of the AMTP concept. The exact design features of the current generation of improved ARTEP documents will be adopted, based upon the results of this field test, and then schools will be expected to begin full implementation of the new ARTEP design concept. Implementation of this concept will involve a substantial increase in the ARTEP development workload.

Under the original ARTEP concept, the analytical work of ARTEP developers was largely complete when tactical doctrine ("how to fight" manuals) and other sources of information had been analyzed through the process of front-end analysis (FEA) to identify the missions and tasks to be addressed by the ARTEP. The final ARTEP document closely resembled the outcome of the FEA process. With the evolution of ARTEP design, ARTEP developers assume the responsibility for analyzing training-relevant features of missions and tasks

and using the results of these analyses to develop training exercises (Drills and STXs) and incorporate training exercises within a descriptive unit training plan.

The evolution of ARTEP greatly expands the ARTEP development audit trail. In effect, the old audit trail involved linking tactical doctrine, FEA results and ARTEP products, in a situation where FEA results and ARTEP were virtually identical. Under the latest ARTEP concept the audit trail involves linking tactical doctrine and FEA results to a variety of ARTEP products. Further, the new audit trail involves linking the various ARTEP products to each other.

The growth in the ARTEP development workload associated with implementing the latest ARTEP design concept comes at a time when many schools are beginning to feel the effects of an ongoing Army Force Modernization effort (i.e., the adoption of new technology and organizational structures). Force modernization, in itself, increases the number of unit types for which ARTEPs must be developed and it increases the frequency with which certain ARTEP documents need to be revised to reflect changes in tactical doctrine. The careful application of computer technology to the ARTEP development process is required to help service schools effectively implement improved ARTEP concepts and deal with changes required by the Force Modernization program.

Mechanism Being Used to Support the Changing Role of ARTEP Developers

ARI and the U.S. Army Training Board (ATB) are developing a Computer-Based ARTEP Production System (CAPS) to support the preparation and revision of AMTP/Drill documents. The portion of the AMTP/Drill development process to be addressed by CAPS begins with the FEA and continues to the point where camera-ready copies are available for printing. In addition, the CAPS concept is intended to support prompt revision of tactical doctrine literature, FEA and ARTEP products in response to changes in tactical doctrine. In brief, a portion of the CAPS database would include tactical doctrine literature and FEA, and the system would guide collective training developers in using the database to develop AMTP/Drill documents.

A preliminary CAPS concept design study indicated that a Relational Database Management System (RDBMS) might serve as an effective core for a CAPS (Bloedoern, Crooks, Saal, Merrill, Meliza and Kahn, in press). A RDBMS stores data in a flexible, tabular fashion and allows information to be extracted from within or across tables using brief, one-line commands.

A RDBMS appears to be particularly well suited to the ARTEP development process, because it can accommodate the complex ARTEP development audit trail in an economical fashion. For example one RDBMS table might contain listings of the references from tactical doctrine for each collective task identified through FEA, and another table might contain a listing of the collective tasks addressed by specific STXs. Through the use of a one-line command, the information within these two tables can be reorganized to provide a table listing the references from tactical doctrine for each STX, as needed.

Application of a RDBMS to AMTP/Drill Preparation

A major goal of applying a RDBMS to AMTP/Drill preparation is to reduce the complexity of work. In general, a RDBMS has the potential to facilitate the complex work involved in preparing AMTP/Drill documents in two ways. First, formal training development guidance and less formal job aids (e.g., "rules of thumb" for making decisions and examples of the application of these rules) might be contained within the database in a manner which relates specific job aids/guidance to appropriate jobs in the ARTEP development process. Second, the analytical capabilities of a RDBMS (i.e., rapid reorganization of data to meet specific information needs) might be used to facilitate complex decisions. The decision processes involved in AMTP/Drill preparation were studied in considerable detail with the goal of reorganizing this process, as necessary, to take advantage of potential applications of a RDBMS to ARTEP development.

Application of a RDBMS to AMTP/Drill Preparation

Work complexity, in this instance, is defined in terms of 1) the number of data elements to be considered and 2) the source of data elements. Certain decisions only require reorganizing information found within tactical doctrine and FEA results. In such cases, the ability of RDBMS to rapidly select and reorganize information can be employed to entirely address decision complexity (i.e., the RDBMS can "make" the decision for the AMTP/Drill developer). Other decisions require consideration of data elements that do not pre-exist in tactical doctrine/FEA. Such data elements are carefully deliberated "judgments" made by AMTP/Drill developers. While reorganization of tactical doctrine/FEA might facilitate human judgments, reorganization does not remove the need for these judgments. Complex decisions of this variety need to be addressed by both "job aids" and information reorganization and retrieval capabilities of a RDBMS.

Three critical patterns become evident when reviewing the entire sequence of decisions to be made by AMTP/Drill developers. First, judgments are required throughout the AMTP/Drill development process. Second, certain judgments serve as input for more than one decision, while other judgments are employed only once. Third, virtually all of the judgments made after the selection of "slices of battle" to be addressed by STXs and Drills are repetitions of judgments made during the selection of STXs/Drills.

Optimum application of a RDBMS to AMTP/Drill document preparation involves minor adjustments in AMTP/Drill development procedures. "Job aids" and the ability of a RDBMS to rapidly select and reorganize information might be employed to facilitate judgmental decisions at the start of the AMTP/Drill development process (i.e., during the selection of STXs and Drills). The results of judgmental decisions relevant to subsequent steps in AMTP/Drill development would then be recorded in the database. RDBMS facilitation of subsequent decisions in the AMTP/Drill development process would be accomplished largely by the ability of a RDBMS to select/reorganize information within the database.

Applications of a RDBMS to AMTP/Drill preparation will be tested and refined by ARI and ATB through the development of a prototype CAPS within the U.S. Army Infantry School. The results of the present effort serve as input for the design of a CAPS database by identifying: judgmental decisions to be addressed by "job aids" embedded within the database; decisions made by collective training developers which, for reasons of efficiency, need to be recorded within the database; data elements (i.e., potential data files) serving as input for key decisions within the AMTP/Drill development process.

References

1. Bloedorn, G., Crooks, W.H., Saal, H., Merrill, D., Meliza, L. and Kahn, O. Concept Study of the: Computer-Aided ARTEP Production System (CAPS). ARI Research Report. 1403, July 1985.

2. Hiller, J.H., Hardy, G.D., and Meliza, L.L. Guideline for Designing Drill Training Packages, ARI Research Product 84-15, January 1984.

3. TRADOC PAM 310-8, Collective Front-End Analysis for Development of the Army Training and Evaluation Program (ARTEP) and a Method for the Development of Drills, 25 September 1981.

4. TRADOC Reg 310-2 (Test-Revised) Development, Preparation and Management of Army Training and Evaluation Program Mission Training Plans (AMTPs) and Drills, 1 May 1985.

5. TRADOC Regulation 310-2, Development, Preparation and Management of Army Training and Evaluation Program (ARTEP) 1 December 1982.

Application of Model Aircrew Training System (MATS) to B-52 Combat Crew Training

Conrad G. Bills, Major, USAF
Instructional Systems Development Division
93d Bombardment Wing
Castle AFB CA 95342-5000

Robert T. Nullmeyer, PhD
Operations Training Division
Air Force Human Resources Laboratory
Williams AFB AZ 85224

INTRODUCTION

Background

The driving force behind the development of the Model Aircrew Training System (MATS) has been the significant advancement in training simulation capabilities with the advent of the full scale weapon system trainer (WST). Major Miller (1978) concluded that the effective use of this technology would require deliberate integration of the WST, to include: (a) identification of training requirements for the B-52 WST, (b) concurrent training accomplishment report (TAPR) for events in the WST, (c) concurrent scheduling function for both WST and aircraft, and (d) Director of Training development of the WST syllabus to insure training requirements are specified. Since the application of WST technology applies across all phases of training, MATS is to be a total training system which integrates training across the full life cycle of the crewmember in the weapon system, incorporating all media, including the aircraft. Prior to MATS, contracted training system development has been primarily focused on ground training (KC-10, C-5).

The 1980 Board of Visitors report on B-52/KC-135 aircrew training management recommended an investigation of "more effective ways to use training devices and simulators to accomplish training objectives and use available flying time for critical tasks." They noted that Computer-Managed Instruction (CMI) methods would pay large dividends in optimizing the interrelated academic/simulation/flying training components of the total training program. The 1982 Board of Visitors concurred with prior recommendations and emphasized the need for automation of training management, evaluation and delivery.

Also in 1982, the USAF Scientific Advisory Board (SAB) concluded that the Military Airlift Command (MAC) training needed to become more efficient by applying currently available training technology. A major problem identified by the SAB was that the current level of combat training is inadequate, but that additional resources are not available to provide the necessary training within the Air Force. They recommended developing a model training program that would embody state-of-the-art training practices and technology to support reallocation of training resources in a manner that more effectively prepares MAC aircrews for their combat mission. They specified the C-130 weapon system due to the availability of advanced aircrew training devices at the present time. They recommended that this program be a model system for

other programs to emulate since the basic problems in C-130 training are common to other weapon systems. A subsequent MAC aircrew training task force concurred with this recommendation.

About the same time as the SAB study, the follow-on operational test and evaluation (FOT&E) was being conducted for the C-130 WST (Nullmeyer & Rockway, 1984). The study was accomplished in three phases corresponding to initial qualification (Phase I), mission qualification (Phase II) and continuation (Phase III) training. Training in the WST generally transferred positively to the aircraft as measured by proficiency ratings and by sorties to criterion. Substantial differences were shown between the performance of students trained in the WST (with visual cues) over students trained in operational flight trainers (OFT, without visual cues). The primary effect of the WST was putting the right person, the student, at the controls when they went to the airplane, rather than the instructor. The recommendations were: (a) establish clear goals for simulator training, (b) train to a set criterion in the simulator, (c) allow simulator training to accommodate individual differences (among instructors as well as students), and (d) integrate simulator training into the overall training system to take advantage of WST training potential.

In response to these problems, many of which exist throughout Air Force training, AFHRL initiated a contract for a model aircrew training system (MATS). This contract focused on the C-130 aircrew training program where lessons learned from the C-130 WST FOT&E could be used along with other state-of-the-art training technology. Attention was directed towards inefficiencies of the event driven program, heavily labor-intensive and aircraft dependent for all stages of training, resulting in some critical areas (e.g., combat tactics training) receiving inadequate attention because available training resources were allocated elsewhere. MATS could also have application to aircrew training for other weapon systems both within MAC and in other major air commands (MAJCOMS).

The MATS contract specified development of an integrated total training system for the full continuum of C-130 aircrew training. The system would provide for training program development, training delivery, training management, evaluation and training analysis, and training support. The system would include: (a) use of advance training concepts and state-of-the-art instructional technology, (b) proficiency-based instruction with advancement based on demonstrated performance, (c) effective use of existing resources and training media, especially the full exploitation of the combat training capabilities of the C-130 WST and most efficient use of available flight time, (d) improved capabilities for more systematic definition of task training requirements, aircrew performance standards, and training media necessary for proficiency-based training and aircrew quality control, (e) use of computer-supported resource scheduling, record keeping, and reporting, (f) provisions for systematically incorporating advances in training technology and program updates, (g) provisions for both internal and external feedback for evaluation and quality control, and (h) provisions to accommodate variable trainee experience levels, learning rates, and class composition. Emphasis was to be placed on computer-based instruction (CBI). Instructor training would need to reflect a new role for more direct student/instructor contact. This system is to serve as a prototype with generalization not only to other MAC programs, but also to Strategic Air Command (SAC) and Tactical Air Command (TAC) aircrew training.

B-52 WST FOT&E

B-52 WST FOT&E was initiated in 1983 on the preproduction unit, but was terminated in June 1983 due to the unreliability of the device. With the acceptance of the production unit, FOT&E was resumed in June 1984. The critical issue was transfer of training from the WST to the aircraft. In addition to the transfer of training study, the plan also included comparisons among WST training options (Table 1), initially four options and subsequently followed by option five and option nine (Table 2).

Table 1

Sequence for Five B-52 WST Training Options

Option 1	Option 2	Option 3	Option 4	Option 5
				WST-0
WST-1	WST-1	WST-1	WST-1	WST-1
WST-2	WST-2	WST-2	WST-2	WST-2
F-1	WST-3	WST-3	WST-3	WST-3
WST-3	WST-4	WST-4	WST-4	WST-4
F-2	WST-5	F-1	F-1	(WST-5)
WST-4	WST-6	F-2	F-2	F-1
F-3	WST-7	F-3	WST-5	WST-6
WST-5	F-1	F-4	F-3	F-2
F-4	F-2	F-5	F-4	F-3
F-5	F-3	F-6	F-5	WST-7
WST-6	F-4	F-7	WST-6	F-4
F-6	F-5	F-8	F-6	F-5
F-7	F-6	F-9	F-7	F-6
F-8	F-7	F-10	F-8	F-7
F-9	F-8	F-11	F-9	WST-8
F-10	F-9	F-12	F-10	F-8
F-11	F-10	F-13	F-11	F-9
WST-7	F-11		WST-7	WST-9
F-12	F-12		F-12	F-10
F-13	F-13		F-13	F-11
				F-12
				F-13

Table 2

Sequence for B-52 WST Training Option Nine

Crew	Pilot	Navigator	Electronic Warefare	Gunner
F-9	Solo			
F-8	SACR 60-4 Checkride			
F-7				
(W-11	Optional)			
F-6				
(W-10	Optional)			
F-5				
F-4	First Integrated Crew Flight			
W-9	Solo (combat sortie)			
W-8	Progress Check			
W-7				
F-3	Part Task Training			
W-6				
W-5				
W-4				
F-2				
W-3				
	F-1	W-2	W-2	W-2
	W-1	W-1	W-1	W-1

By August 1984 initial results of B-52 WST FOT&E were available concerning training using the first four options. B-52 WST was showing good reliability, availability, and maintainability. Instructors liked it, particularly for the benefits in improved crew coordination. Transfer of training seemed to be positive. The need for computer assisted scheduling was evident as well as the need to add WST instructor options for individualizing instruction. Additional days were added for WST mission planning/critique. Development of training option five was initiated for a better building block approach in WST training.

In November 1984, the tryout of training option five began. The number of WST missions was increased from seven to ten to include a part task training mission for pilots only. The first two WST missions were abbreviated, one training air refueling only and the other low level only. Subsequent full missions continued development of crew coordination, provided instructor options to meet individual student needs, and then gave opportunities for enhancing crew coordination. Training option five was scheduled for four of the eight crew lines due to limitations of only one operational B-52 WST and insufficient console operator manning.

In January 1985, B-52 WST FOT&E results for the first four training options were reported. In October 1985, training option nine results were added. Training in the WST generally transferred positively

to the aircraft. Time to initial proficiency as measured by instructors on B-52 student progress sheets was significantly shorter for initial qualification students with WST training than for students without WST training (Tables 3, 4, 5 & 6). By the end of training there was little difference between the proficiency of the two groups as measured by standardization/evaluation personnel on SACR 60-4 checkrides. All instructors reported gains in WST training, particularly crew coordination. They indicated that student performance during the first aircraft sortie was comparable to that of students on sortie two or three of the non-WST syllabus. After seven or more front-loaded WST missions, student pilots were proficient enough to fly in the seat low level the first integrated aircraft sortie. Safety was excellent with no mishaps reported.

Table 3

Comparison of Number of Trials to Rating of 3.0/4.0 on B-52 Progress Sheet for Initial Student Crews with Weapon System Trainer (WST) Syllabus and without WST Syllabus

Position	N	Rating	Q*	p**	JIC***	p**
Copilot	38	3.0	.21	NS	.31	.05
		4.0	.30	.05	.46	.005
Navigator	35	3.0	.04	NS	.25	NS
		4.0	.05	NS	.39	.025
Electronic Warfare Officer	29	3.0	.42	.025	.09	NS
		4.0	.28	NS	.64	.005

*Coefficient of Colligation (consistency of effect)
**Fisher's Table of Correlation Coefficients
***Jenkins Index of Covariation (magnitude of effect)

Table 4

Comparison of Number of Sorties to Rating of 3.0/4.0 on B-52 Progress Sheet for Initial Students Crew with Weapon System Trainer (WST) Syllabus and without WST Syllabus

Position	N	Rating	Q*	p**	JIC***	p**
Copilot	38	3.0	.24	NS	.40	.01
		4.0	.43	.005	.40	.01
Navigator	35	3.0	.29	.05	.29	.05
		4.0	.24	NS	.61	.005
Electronic Warfare Officer	29	3.0	.59	.005	.40	.025
		4.0	.59	.005	.69	.005

Table 5

Comparison of Number of Trials to Rating of 3.0/4.0 on B-52 Progress Sheet for Training Option Nine Student Crews with Weapon System Trainer (WST) Syllabus and without WST Syllabus

Position	N	Rating	Q*	p**	JIC***	p**
Copilot	22	3.0	.56	.005	.40	.05
		4.0	.38	.05	.64	.005
Navigator	21	3.0	.53	.01	.38	.05
		4.0	.91	.005	.44	.025
Electronic Warfare Officer	19	3.0	.49	.025	.11	NS
		4.0	.41	.05	.67	.005

Table 6

Comparison of Number of Sorties to Rating of 3.0/4.0 on B-52 Progress Sheet for Training Option Nine Students Crew with Weapon System Trainer (WST) Syllabus and without WST Syllabus

Position	N	Rating	Q*	p**	JIC***	p**
Copilot	22	3.0	.83	.005	.43	.025
		4.0	.98	.005	.57	.005
Navigator	21	3.0	.52	.01	.33	NS
		4.0	.99	.005	.69	.005
Electronic Warfare Officer	19	3.0	.15	NS	.50	.025
		4.0	.86	.005	.77	.005

Generally for each B-52 WST training option, as the number of WST missions increased, the number of aircraft flights decreased without degradation of final proficiency. This was particularly true going to training option five and then again to training option nine. Graduation point was dropped from fifteen to thirteen aircraft sorties and then down to nine.

Training option nine. In February 1985, training option nine was designed to bring each crew member to a desired level of proficiency and pass a required progress check in the WST before going on to integrated flight

training in the aircraft. This training option reduced aircraft flying sorties by approaching proficiency based training. All aircrew training was rehearsed in the WST or another intermediate level training device prior to accomplishment in flight. For events in which the WST was inadequate, the aircraft supported "part task training." Standardization/evaluation personnel performed the progress check in the WST and flew with the crew on the first sortie following the progress check to assess the level of proficiency attained at that stage in training. Again, by the SACR 60-4 check there was no appreciable difference in proficiency between student crews trained with WST and those trained without WST. This was the first training option for which a list of training events was approved for creditation in the WST.

Though training option nine was a leap ahead toward effective use of the B-52 WST, the results of the B-52 WST FOT&E also confirmed the C-130 WST FOT&E results that lessons learned from the C-130 MATS study are needed to maximize the capabilities of WST technology. While monitoring acquisition of the C-130 aircrew training system (ATS), the B-52 program needs to develop improved methods for instructor training and for student preparation in order to take advantage of the enhanced capabilities of the WST. Also, the B-52 program needs continued effort in developing a computer based training system with CMI and CAI. Computer based training (CBT) applications will be particularly computer assisted scheduling and for procedures reinforcement, system simulation, and part task training. This will require a review of the entire training system from the MATS perspective.

B-52 APPLICATION OF MATS

Similarity of training.

C-130 aircrew training is sufficiently similar to B-52 aircrew training so that there can be direct spin-offs from C-130 MATS. Both training programs have the three phases for aircrew training, i.e., initial qualification, mission qualification, and continuation.

Initial findings of the C-130 MATS phase I (Fishburne, Williams, Chatt, & Spears, 1984) showed that though instructional systems development (ISD) methodology was purported, the program lacked systems design. Most of the training resources were for the formal school. Instructional strategy was transfer as a product of learning--learn then apply. Instruction was adapted to the "average" trainee. Simulation was not a key element in aircrew training. Flying instruction was primarily in-flight hours. Continuation training was deficient and piecemeal. In comparison with B-52 training, these elements were quite similar.

The C-130 MATS phase I report indicated training program needs which also apply to B-52 training: (a) unify the management of aircrew training, (b) individualize instruction, (c) take maximum advantage of computer-based training technology, and (d) apply modern concepts of learning.

Increased productivity initiative.

Though the results are not all in from MATS, the cross feed between B-52 and C-130 aircrew training programs has given opportunity to apply lessons learned in developing the B-52 Combat Crew Training School (CCTS) increased

productivity initiative for FY86. B-52 aircrew training has shifted toward the more desired 1:1 student/instructor ratio and yet still increasing student output. A key component will be a firm progress check in the WST with pass/fail criterion and additional training required until desired proficiency is achieved before moving on to integrated aircraft flight activity. Essential elements of a computer managed instruction system will be instituted to keep a pulse on the effects resulting from implementation of the FY86 training plan.

SAC MATS working group.

Headquarters SAC and Human Resources Laboratory, Williams AFB, AZ, are combining forces to review MATS concepts as they apply to SAC aircrew training. They will identify lessons learned from other training programs as well as optimal application of computer based training (CBT). Their goal is to develop a generic SAC model with organizational structure, functions of components in the structure, the information/feedback interfaces of all program elements, the SAC/contractor interface, and the identification of changes needed to Air Force and SAC directives. MATS concepts will initially be applied to the C-130 ATS. As SAC ATS is developed, continued cross feed will facilitate ready application of lessons learned along the way.

REFERENCES

B-52 weapon system trainer follow-on operational test and evaluation, final report. Castle AFB, CA: Instructional Systems Development Division, October 1985.

Board of Visitors report on B-52 and KC-135 aircrew training management. Castle AFB, CA: Instructional Systems Development Division, February 1980.

Board of Visitors report on B-52/KC-135 aircrew training. Castle AFB, CA: Instructional Systems Development Division, May 1982.

Fishburne, R. P., Jr., Williams, K. R., Chatt, J. A., & Spears, W. D. Design specification development for the C-130 model aircrew training system: phase I report. Seville Training Systems technical report, Irving, TX, January 1985.

Jenkins, W. O. & Hatcher, N. C. The design of behavioral experiments. Auburn University at Montgomery, AL (unpublished manuscript), 1976.

Miller, S. J. Integration of the B-52 weapon system trainer in the combat crew training school program. Air Command and Staff College research report, Air University, Maxwell AFB, AL, May 1978.

Nullmeyer, R. T., & Rockway, M. R. Effectiveness of the C-130 weapon system trainer for tactical aircrew training. Paper presented at the Interservice/Industry Training Equipment Conference, Washington, D.C., October 1984.

Williams, K. R., Marcantonio, A. W., & Fishburne, R. P., Jr. C-130 model aircrew training system: functional design specification. Seville Training Systems draft technical report, Irving, TX, July 1985.

Sensitivities of Speeded Subtests

Toni G. Wegner
Malcolm James Ree
Air Force Human Resources Laboratory

Most aptitude tests are classified as power tests or speeded tests. Power and speeded tests differ in their emphasis on cognitive capabilities versus speed of processing. The purpose of this paper is to demonstrate that scores on speeded tests can be influenced by a variety of factors, and to discuss the implications of this in the use of speeded tests.

Power tests are designed to tap a designated cognitive capability and are structured to allow a majority of subjects to complete the test. In contrast, purely speeded tests are designed to measure speed without measuring deep cognitive capabilities; these tests consist of fairly easy items so that most of the people who attempt an item will answer it correctly, but few people will complete all items. There is a consensus within the testing literature that standardized testing requires strict control over the test administration procedures, especially such things as controlling the testing environment, time limits, and instructions (see, e.g., Anastasi, 1976; Clemans, 1971). Relatively little research, however, has addressed the deviations of administration procedures, especially as they relate to differential impact on power versus speeded tests.

The Armed Services Vocational Aptitude Battery (ASVAB) is an aptitude battery used for the selection and classification of all military enlisted personnel. Since 1980, the ASVAB has consisted of ten tests--eight power and two speeded. The power tests measure such things as verbal, mathematical, and technical information; the speeded tests, Numerical Operations (NO) and Coding Speed (CS), measure perceptual-psychomotor ability. In research and operational use of the ASVAB, speeded subtest scores have been shown to vary as a function of a variety of factors. These factors, then, produce error variability that is not found to the same extent in the power tests. Data are available to identify three factors that can produce fluctuations of speeded test scores: answer sheet format, test booklet format, and practice.

Answer sheet format effects first became evident after the ASVAB scores obtained from a carefully collected sample of American youth (McWilliams, 1980) were found to deviate from scores of military personnel only on the speeded tests. An initial study was conducted using Air Force basic recruits (Earles, Giuliano, Ree, & Valentine, 1985) to test the hypothesis that differences in the kinds of grids on the answer sheets used by the American youth sample versus military personnel could account for the differences in scores. A large-scale study involving applicants for all the armed services provided confirmation that the answer sheet differences account for the difference in performance between the two groups (Wegner & Ree, 1985).

Further evidence of answer sheet format effects came from another study with Air Force basic recruits. In this study, a third type of answer sheet, one used solely for ASVAB research purposes, was found to produce differences in speeded test scores (Wegner & Ree, 1984). In all three of the above

studies, subjects were administered identical test items using one of two answer sheets in an equivalent groups design. In no case were score differences found between the two groups on any of the power tests. The speeded test score results for the three studies are presented in Table 1. The answer sheets used differed in size, shape, and spacing of the response grids; this was enough to make a significant difference in scores on both the Numerical Operations and Coding Speed tests even at the raw score level.

Table 1. Answer Sheet Format Comparisons

Study	Test	Operational Answer Sheet	Other Answer Sheet	Difference
Earles, et al. (1983)(N = 512)	NO	41.22	37.72	3.50*
	CS	52.33	51.01	1.32
Wegner & Ree (1984)(N = 502)	NO	39.71	34.67	5.04*
	CS	50.53	47.70	2.83*
Wegner & Ree (1985)(N = 8598)	NO	35.83	32.64	3.19*
	CS	46.93	45.59	1.34*

*Significant at the .05 level.

Note. Numbers are cell means. NO maximum is 50; CS maximum is 84.

Like answer sheet format effects, test booklet format effects occur as part of the test materials used. In the case of booklet format effects, differences between booklets were very subtle. Small differences in such things as type font, character pitch, and spacing produced significant variations in speeded test scores. Booklet effects were suspected when systematic differences were found in a single testing period between a form of the ASVAB and a scrambled version of the same form. The original and scrambled forms contained identical items in a slightly different order. The differences in scores were found consistently at monthly testing intervals. Speeded test scores for two months of testing, six months apart, are presented in Table 2 for the original form and the scrambled version of the same form. It is interesting to note that booklet differences did not impact both speeded tests uniformly. The same format that benefited Numerical Operations scores in one type of booklet proved to be detrimental to Coding Speed scores. No substantial systematic differences were found on the power tests.

Table 2. Test Booklet Format Effects

Test		Original	Scrambled	
	Time 1	(n = 9119)	(n = 7651)	Difference
NO		38.64	37.67	.97
CS		46.60	50.99	-2.39
	Time 2	(n = 8267)	(n = 7211)	Difference
NO		38.07	37.07	1.00
CS		47.39	50.13	-2.74

Note. Numbers are cell means. NO maximum is 50; CS maximum is 84.

Practice effects, unlike format effects, depend on the test-taker rather than characteristics of the test. If a person takes a test more than once, it is implicit in the theory of testing that his or her score will not be identical each time. This is because a given score is considered to be an estimate of the person's true score, and is expected to vary within a range that specifies the accuracy with which the test can measure. Given this logic, it is reasonable that retests of the ASVAB sometimes show increments and sometimes show decrements in scores on power tests. The same does not hold for speeded tests. When all other conditions are held constant (e.g., motivation), scores on speeded tests are most likely to show uniform increments on subsequent tests. These effects can be attributed to practice, and reveal a condition under which scores on speeded tests can be enhanced relative to scores on power tests in a testing situation. Table 3 shows speeded test results for a single form of the test--one group had previously taken a similar form of the test (and thus had practice), the other group had not. These data were collected as part of the operational calibration of the most recently implemented versions of the ASVAB (Forms 11/12/13), with each score based on data from about 25,000 applicants.

Table 3. Practice Effects

Test	Practice	No Practice	Difference
NO	39.73	38.61	1.12
CS	52.25	48.49	3.76

Note. Numbers are cell means. NO maximum is 50; CS maximum is 84.

The three factors described above clearly do not exhaust the conditions under which speeded tests show sensitivity. Obvious problems can result, for example, when timing discrepancies occur. Consider the Numerical Operations test in which subjects are given three minutes to attempt 50 items. With subjects answering 35 items correctly on the average, it takes approximately five seconds to answer each item. Sloppy timing of even a few seconds can have significant effects.

Experience with the above issues has shown that the way to deal with the sensitivity of speeded tests depends on which of the factors is in question. If score comparisons are to be made at all on the basis of a test (which is a major reason to administer standardized tests), both answer sheet format and test booklet format need to be considered. It is preferable that tests be administered using identical booklets and answer sheets. If this is impossible (for example, due to obsolete answer sheets or the use of existing data), an adjustment should be developed to make the scores comparable (see Wegner & Ree, 1985, for one way this can be done). Practice effects, if they cannot be controlled, should be noted in the interpretation of test scores. In the use of tests likely to be affected by practice, one might consider incorporating practice into the instructions for the test so that outside practice effects are minimized.

In summary, speeded tests are sensitive to a variety of internal and external factors that do not similarly affect power tests. These effects on speeded test scores can be minimized by avoiding, controlling, or taking into consideration these factors.

References

Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Clemans, W. V. (1971). Test administration. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed) (pp. 188-201). Washington, DC: American Council on Education.

Earles, J. A., Giuliano, T., Ree, M. J., & Valentine, L. D., Jr. (1983). The 1980 youth population: An investigation of speeded subtests. Unpublished manuscript, Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

McWilliams, H. A. (1980). Profile of American youth: Field report. Chicago, IL: National Opinion Research Center.

Wegner, T. G., and Ree, M. J. (1984). Comparison of ASVAB scores on operational versus research answer sheets. Unpublished manuscript, Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Wegner, T. G., and Ree, M. J. (1985). Armed Services Vocational Aptitude Battery: Correcting the speeded subtests for the 1980 youth population (AFHRL-TR-85-14). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Initial Operational Test and Evaluation of Armed Service Vocational Aptitude Battery (ASVAB) Forms 11, 12, and 13: Data Quality Analysis

John R. Welsh
Toni G. Wegner
Air Force Human Resources Laboratory

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple test battery that is used by the Department of Defense (DOD) and the United States Armed Forces to select and classify applicants for military service. Operational versions of the test are replaced periodically. As new versions are implemented, tables must be generated to convert scores from the new tables to the metric of the reference population. Temporary tables are developed before the test is used operationally through an operational calibration. Once the versions are put in operational use (using the temporary tables), an Initial Operational Test and Evaluation (IOT&E) is conducted to finalize the tables. This paper describes data quality issues surrounding the IOT&E of ASVAB Forms 11, 12, and 13.

Since October 1980, the ASVAB has been composed of ten subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). Combinations of two or more subtests are used by the services to determine the specialty and occupational classification of applicants. The Armed Forces Qualification Test (AFQT) is a composite used by all the services for determining enlistment qualification of applicants; it consists of AR, WK, PC, and half-weighted NO.

Operational versions of the ASVAB are replaced by DOD and the joint services every three years to update obsolete items, take advantage of advances in the field of psychometrics, and minimize exposure of the battery. Replacement of the battery generates the requirement to equate new forms of the test to a reference test. The accurate establishment of the new tests on the score scale of the reference test is essential for meaningful year-to-year comparison of the distribution of abilities of applicants and recruits, as well as providing consistent meaning of scores used to make classification decisions.

In October 1984, the reference population was changed from a 1944 population based on men under arms to a representative sample of 1980 American youth (males and females, ages 18-23). This change was made to provide manpower and personnel planners with a more relevant normative base from which to make manpower decisions. The representative sample of 1980 American youth was tested on ASVAB Form 8a, which was implemented for operational testing in October 1980.

The most recent versions of the ASVAB are Forms 11/12/13. These batteries were designed to be parallel to ASVAB Form 8a. An operational calibration of ASVAB Forms 11/12/13 was conducted in May and June 1982 to develop tables to convert scores to the 1980 American youth reference

population (based on Form 8a scores). Forms 11/12/13 were scheduled to be implemented in October 1983. Subsequent issues surrounding the sensitivity of the ASVAB speeded subtests (NO and CS) to differences in answer sheet format caused a one-year delay in the planned implementation of the new forms (See Ree, Welsh, Wegner, & Earles, 1985). Corrections for answer sheet format effects were made (Wegner & Ree, 1985), and Forms 11/12/13 were implemented in October 1984 on the 1980 metric.

The IOT&E of ASVAB Forms 11/12/13 was conducted in Oct and Nov 84, immediately following implementation. Data were collected at all Military Entrance Processing Stations (MEPS) on each of the two versions of the three new forms (i.e., 11a, 11b, 12a, 12b, 13a, and 13b) and the reference test (Form 8a, labeled 13c for the IOT&E study) in an equivalent groups design. The purpose of the IOT&E was to establish the defining relationship of the six new ASVAB forms to the reference test under operational conditions. The data were to be used to check the operational calibration of the new forms, and to generate the final operational conversion tables.

All applicants who tested for the services between 1 Oct 84 and 30 Nov 84 had their answer sheets scanned at the Air Force Human Resources Laboratory; the results were provided to a contractor to be analyzed. During the data collection phase of the IOT&E, anomalies in mean AFQT raw scores of applicants taking the six new forms and the reference test were noted. This report will describe the data quality issues surrounding the IOT&E in the investigation of these anomalies. A more complete description of the development of the newly implemented ASVAB forms is contained in Prestwood, Vale, Massey, and Welsh (1985).

The Problem

Anomalies were noted when the AFQT means for the six new forms and the reference test based on the operational calibration (OPCAL) were compared with the same means based on IOT&E data. In the OPCAL, all seven versions were administered to service recruits; 11a and 8a/13c were also administered to service applicants. In the IOT&E, all data were obtained from service applicants. Whereas the 13c AFQT mean is about the same as the means for the other versions in the OPCAL, it is clearly higher than the means of the other versions in the IOT&E. These data are presented in Table 1.

Table 1. Mean Raw AFQT Scores by Test Form

Form	OPCAL		IOT&E
	Applicants	Recruits	Applicants
11a	72.26	75.21	73.63
11b		75.31	73.42
12a		73.20	71.51
12b		74.90	73.49
13a		75.11	73.82
13b		75.20	73.49
8a/13c	72.17	75.02	74.49

Because the IOT&E is the defining relationship between the new forms and the reference test, these data are valid as long as there is no reason to suspect the data are faulty. The anomalies between the AFQT means from the OPCAL and IOT&E raised suspicion and warranted investigation of the anomalies to verify the appropriateness of the use of the IOT&E data. The data quality analyses investigated four possible explanations.

First, it was possible that the groups taking the seven batteries were not equal in ability and that, by chance, the group that took Form 13c was slightly smarter. The second possibility was that the reference test was compromised to some extent, especially since Form 8a had been used operationally for two years. Evidence for nonequivalent groups or compromise in the IOT&E would invalidate these data as the defining relationship between the new forms and the reference test. The third possibility existed that there might have been a system-wide scoring error (key-error or conversion table error) in the automated MEPS scoring system. An error of this kind would be correctable. Finally, it was possible that format or printing differences between OPCAL and the IOT&E caused the observed mean score differences. The impact of these would depend on the nature of the differences.

Analyses

The analyses led to the elimination of the first three possibilities. The equivalence of the groups was examined by comparing the raw score performance of all seven groups on the General Technical composite. This composite contains the same subtests as the AFQT, with the exception of NO. Table 1 shows the groups were generally quite equal on ability. This also suggests that the raw score elevation of AFQT on Form 13c was due to NO.

Table 2. Mean Raw General Technical Composite Scores by Test Form

Form	11a	11b	12a	12b	13a	13b	13c
GT	55.01	54.34	54.29	55.72	55.33	55.48	55.32

Note: Data are based on 18,000 plus applicants per form.

The second explanation was ruled out because there was independent evidence that compromise was at a low level at the time of the IOT&E (Sims, Truss, & Curia, 1985). A spot check of the MEPS scoring system revealed no error in the test scoring or in the use of appropriate conversion tables.

Given that the raw score AFQT differences could be traced to NO, format or printing differences were the most plausible explanation for the elevation of the 13c mean. This is consistent with other research (Wegner & Ree, 1985) showing speeded subtests to be sensitive to answer sheet format effects. Inspection of the IOT&E test booklets for the seven versions of the ASVAB revealed slight differences between the style used to print Form 8a/13c and that used to print the other six tests (which were identical). These differences can be characterized by the closer distance between letters and

numbers on the reference test. The type facing was also somewhat bolder. It should be noted that the format used for Form 8a/13c was identical to that used to collect data on the reference population.

The spacing on the NO subtest was visually more compact on Form 8a/13c, suggesting the format could account for Form 13c being slightly easier or faster. A study was conducted with Air Force basic recruits to test the hypothesis that format effects could account for the originally observed raw score anomalies. Identical NO items were printed in two formats: the 8a/13c format and the format used for the other six versions. These were randomly administered to two groups, each containing about 120 recruits. The mean NO raw scores were 41.4 for the group with the Form 8a/13c format and 40.1 for the group with the Forms 11/12/13 format. The results indicate that the Form 8a/13c format used in the IOT&E was slightly faster than the format of the other versions. Because of ceiling effects on NO with Air Force recruits, the magnitude of the difference is smaller than would be obtained with a broader range applicant sample; however, this difference still accounts for most of the mean AFQT difference between Form 8a/13c and the other six versions in the IOT&E (see Table 1).

Conclusions

It was concluded from these analyses that the difference in mean AFQT raw scores between Form 8a/13c and Forms 11/12/13 was probably due to the slightly "faster" print format used for the reference test. For comparison purposes, the print formats for the seven versions used in the OPCAL were also inspected. These booklets used a format that was different than either of those used in the IOT&E, but all seven versions were printed with the same format. This explains why the pattern of AFQT means was different for the OPCAL and the IOT&E.

Because the format used for Form 8a/13c in the IOT&E is the same as that used for the 1980 American youth reference population, and there is no reason to suspect the IOT&E data set is faulty, the IOT&E data are appropriate to use for the development of final conversion tables. These data define the relationship between Forms 11/12/13 and the reference test.

References

Prestwood, J.S., Vale, C.D., Massey, R.H., & Welsh, J.R. Armed Services Vocational Aptitude Battery: Development of forms 11, 12, and 13 (AFHRL-TR-85-16). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Ree, M.J., Welsh, J.R., Wegner, T.G. & Earles, J.A. (1985). The equating and implementation of the Armed Services Vocational Aptitude Battery (ASVABs 11, 12, and 13) in the 1980 youth population (AFHRL-TP-85-21). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sims, W. H., Truss, A. R., Curia, M. (1985). Extent of cheating on ASVAB (CNA 84-117707). Alexandria, VA: Center for Naval Analyses.

Wegner, T. G. & Ree, M. J. (1985). Armed Services Vocational Aptitude Battery: Correcting the speeded subtests for the 1980 youth population (AFHRL-TR-85-14). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

DEVELOPMENT OF AN INTEGRATED PILOT SELECTION SYSTEM

Jeffrey E. Kantor
Air Force Human Resources Laboratory

INTRODUCTION

The Air Force Human Resources Laboratory (AFHRL) is conducting a multi-year research and development (R&D) program to improve procedures for selecting candidates for USAF Undergraduate Pilot Training (UPT). The principal goal of selection procedures for UPT is to screen out those candidates with low chances of completing training and/or becoming successful operational pilots. This is more important now than ever because the costs of each UPT eliminee have risen dramatically (approximately $64,000) and the operational mission in which USAF aircrews are employed has increased in complexity and difficulty. Therefore, improvements in identifying candidates who have the requisite abilities for success will both reduce the wasted costs associated with eliminations and help ensure that operational pilots are capable of meeting the rigorous demands of today's military aerospace environment.

The first step in this R&D program was to review the current pilot candidate selection procedures. Since 1955, the principal components used in USAF pilot selection have included a physiological fitness examination; the paper and pencil Air Force Officer Qualifying Test (AFOQT); certain personal factors such as age and the type of college degree obtained by the candidate; and actual flying experience, assessed either by the possession of a civilian pilot's license or performance in a short Air Force light aircraft flying program given to promising candidates without a license. While these components do significantly predict success in UPT, literature reviews (Imhoff & Levine, 1981) and discussions with flight training and operational personnel indicated that additional candidate testing in the areas of psychomotor (hand-eye coordination) performance, cognitive abilities such as information processing speeds, and personality factors would have considerable potential for making improvements to the current system. To accomplish testing in these areas, the Basic Attributes Tests (BAT) were developed.

The BAT is a totally automated testing system designed around a commercially available supermicro-computer (Alcyon). The BAT system encompasses all the necessary hardware and software for high speed graphics, millisecond interval response timing, multiple joystick control, and keypad response entry. Software sub-routines in the BAT provide internal diagnostic checks, perform auto-recoveries in the event of systems failures, control test presentations, score test performance, conduct immediate data audits and checks for out-of-range responses and perform data file handling chores. The current version of the BAT has 15 subtests which are being evaluated for aircrew selection. These include measures of perceptual and mental encoding speed, mental rotational abilities, decision making speed, time sharing ability, figure-ground separational ability, risk taking, decisiveness, personality factors, and two tests of psychomotor ability. While the majority of these tests are in the early stages of evaluation, the psychomotor tests have been extensively analyzed and their usefulness in the effort to improve the selection of Air Force pilot candidates is the main topic of this paper.

CLEARED FOR PUBLIC RELEASE
Office of Public Affairs

METHOD

Predictive Validation

The two psychomotor tests were administered to 1725 subjects prior to their entry into UPT during fiscal years 79 through 82. These subjects had been selected by the normal system and their psychomotor scores were held in confidence until after they either completed or were eliminated from pilot training. Air Force UPT is a 49 week course with approximately 175 flying hours conducted in the T-37 and T-38 jet training aircraft. Following training termination, graduate or elimination status was obtained on these subjects and used as the criterion for the predictive validation of the psychomotor tests.

Concurrent Validation

As an additional validation of the psychomotor tests, test units were sent to Williams AFB, Arizona to collect psychomotor scores on 95 graduating UPT students. For this group, the results of the Advanced Training Recommendation Board (ATRB) were obtained as criteria. The ATRB is convened towards the end of UPT and consists of instructor pilots, training squadron commanders and wing staff. The purpose of the ATRB is to decide which of the graduating students should receive a Fighter-Attack-Reconnaissance (FAR) follow-on assignment recommendation and which should receive a Tanker-Transport-Bomber assignment recommendation. Only the better students receive a FAR recommendation and the ATRB decisions were used as the criterion for the concurrent validation of the psychomotor tests.

Psychomotor Test Description

The first of the two tests-TWO HAND COORDINATION-was a pursuit tracking task in which a triangular target moving in a circle had to be tracked with a cross-shaped cursor. The movement of the cursor was controlled by two joysticks. One joystick controlled the left-right axis (X1) movement of the cursor while the other joystick controlled the up-down axis (Y1) movement. The second test-COMPLEX COORDINATION-was a compensatory tracking task in which the subject was required to keep a cursor as close as possible to the intersection of a vertical and horizontal line while also keeping a short bar of light as close as possible to the vertical line. Test generation software included a biasing function which increased the difficulty of this task. The movement of the cursor in the left-right axis (X2) and up-down axis (Y2) was controlled by one joystick while the movement of the short bar of light in the left-right axis (Z2) was controlled by foot operated rudder pedals (later versions of this test, developed after 1983, have eliminated the rudder pedals and use both joysticks for the second test as well).

For both tests, scores are obtained by accumulating the absolute displacements from the cursor to the target point and, for the second test, from the bar of light to the vertical line. Each test has a three minute practice period followed by five minutes of scored performance. Five scores are produced, one of each of the control axes (X1, Y1, X2, Y2, Z2). Since these scores reflect tracking error, lower test scores indicate better performance.

RESULTS

Predictive Validation

For the psychomotor tests to be useful, at least some of the five scores must significantly differentiate between candidates who graduated and those who eliminated from UPT. In addition, it is reasonable to expect some differences among the scores for different types of eliminees. Reasons for elimination from UPT are flying deficiency, academic deficiency, medical, self-initiated, and fatality. If psychomotor ability, as measured by these tests, is related to flying aptitude, then the greatest differences in test scores should be found between the graduates and those who eliminated for flying training deficiency reasons.

These hypotheses were evaluated by comparing the means of the psychomotor scores among three categories--UPT graduates, all UPT eliminees, and UPT flying deficiency eliminees. The mean psychomotor scores for these groups and the probabilities of the differences among these means occurring by chance alone is presented in Table 1. All five scores showed significant ($p \leq .001$) differences in the expected direction between graduates and either category of elimination. In addition, for all three of the COMPLEX COORDINATION scores, the means of the flying deficiency eliminees were significantly ($p \leq .01$) worse than those of the eliminees for all other reasons. These results indicate that both psychomotor tests can identify eliminees but that the COMPLEX COORDINATION test may be a better test to identify those candidates who will eliminate for flying deficiency. These results validate the use of the psychomotor scores as predictors of success in USAF UPT.

TABLE 1

	Psychomotor Scores				
	X_1	Y_1	X_2	Y_2	Z_1
Means by UPT Outcome*					
Graduate Means	14315	16341	3559	2858	4725
All Elim Means	15829	17621	4936	4173	6678
Flying Deficiency (FD) Elim Means	16302	18007	5593	4702	7580
Probabilities of Psychomotor Score Differences Occurring by Chance Alone					
Grads vs All Elims	.001	.001	.001	.001	.001
Grads vs FD Elims	.001	.001	.001	.001	.001
FD vs Other Elims	.103	.182	.006	.010	.005

*Psychomotor scores reflect errors and therefore, lower scores mean better test performance.

Concurrent Validity

The psychomotor testing of UPT students nearing graduation at Williams AFB permitted the concurrent validation of the tests against the criterion of the ATRB decision. Only the better students received a FAR recommendation. It was found that the FAR students had significantly ($p \leq .01$) better scores than the TTB students on two of the five psychomotor measures (a third score was significant at $p \leq .05$). Also, using multiple linear regression, the five psychomotor scores produced a significant ($p \leq .01$) multiple correlation against the FAR/TTB criteria ($R = .43$). These results show that, in addition to identifying candidates with low probabilities of UPT graduation, psychomotor scores also relate to superior performance in UPT. Taken altogether, these results indicate that the quality of UPT students can be improved with psychomotor screening.

Development of a Psychomotor Screening Equation

To obtain the maximum prediction accuracy from the psychomotor scores, a weighted equation or linear model was developed to predict UPT outcome. This equation provides a screening score based on psychomotor ability. Because the criterion was dichotomous and was coded 0 for eliminees and 1 for graduates, the screening score can be roughly interpreted as a probability of success in UPT. To determine the weights for the model, multiple linear regression was used with all five psychomotor scores as predictors and UPT outcome as the criterion variable.

While all five psychomotor scores were significant predictors of UPT outcome, the scores within each test were highly interrelated. Accordingly, the most useful linear model would contain the fewest psychomotor scores which still accounted for as much of the criterion variance as all five scores together. However, that does not mean that the tests could be changed to present only the most predictive axes because that would change the nature of the tasks in the tests. After several model comparisons using the F-ratio, the final screening equation ($R = .196$) contained only the X_1 and Y_2 scores. Using this equation, psychomotor screening scores were computed for the 1725 subjects in the predictive validation analysis and the mean screening score for those who graduated (79.0) was found to be significantly ($p \leq .001$) less from those who eliminated (75.1).

The practical usefulness of this psychomotor screening model was assessed using hit/miss tables produced for the 1725 case sample, comparing the predicted UPT success scores from the final regression equation versus actual UPT outcomes. To illustrate the effectiveness of the psychomotor screening model, two points have been chosen from the hit/miss tables. These two points represent the screening effectiveness of the psychomotor scores had they been used to rank order the 1725 cases and had only the best 90% and 80% of them been chosen. These examples are given in Table 2. For example, using the 10th percentile cut score (taking only the best 90% of the subjects) on this sample would have resulted in correctly rejecting 19.6% of all the eliminees while incorrectly rejecting only 8.0% of the graduates.

Development of an Integrated Pilot Selection System. While this psychomotor-based system appears useful, a potentially more effective approach would consider all valid screening information simultaneously. The next step in the research was to determine how to integrate the psychomotor information into the current selection process. Again, a multiple regression analysis framework was used. A subsample of 268 UPT cases was identified for whom data were available on age, AFOQT scores and performance in the Flight Screening Program (FSP). The FSP is a 14-hour, light-aircraft flying course used to screen applicants on basic flight skills. These are the major components used for selection of non-Air Force Academy pilots candidates. Through regression techniques, an integrated model was developed which captured the unique predictive information available from the experimental psychomotor tests and operational selection factors. The experimental psychomotor tests and every other source of information contributed unique prediction in a final regression model which had a multiple $R = 0.45$ ($p \leq .001$). Another hit/miss analysis was conducted using this integrated model. The effectiveness of the integrated approach, when used to select the best 90% and 80% of the 268-case sample, is also illustrated as part of Table 2. The results were quite impressive. Had the integrated system been used to select the best 90% (a 10 percentile cut score), then 28.9% of all the UPT eliminees in the sample would have been correctly rejected while only incorrectly rejecting 2.6% of the graduates. The integrated approach would have done substantially better than the psychomotor only system.

TABLE 2. Examples of Screening System Effectiveness

	10th Percentile Cut	20th Percentile Cut
Percent of All Elims Rejected		
Psychomotor Only Model	19.6%	31.8%
Integrated Model	28.9%	47.4%
Percent of All Grads Rejected		
Psychomotor Only Model	8.0%	15.7%
Integrated Model	2.6%	10.4%

DISCUSSION

The results of the analyses conducted demonstrate that the two BAT psychomotor tests produced scores which were valid predictors of UPT performance. Candidates who graduated from UPT were differentiated by their scores from those who eliminated. Also, superior UPT students (FAR recommended) were differentiated from weaker students. This differentiation of pilot candidates could be used through the implementation of the psychomotor only screening model, however, a much more effective use was found through the development of the integrated pilot selection system. The integrated system made use of the valid information available from the current operational selection system as well as the psychomotor test scores. Because of the potential improvements in selection accuracy inherent in the integrated

system, in January 1985, the Air Force Air Training Command decided to implement the integrated selection system following an operational trial which has already begun.

Ongoing R&D is continuing with the other BAT subtests. As these tests become validated, their use in the integrated system approach will be evaluated. If it is found that they significantly add to the predictive accuracy of the current integrated system, they will be included in a fashion similar to the psychomotor tests. Because all BAT subtests can be administered on the same computerized testing system, their inclusion in the integrated selection system will entail software changes only. In this regard, the Air Force UPT selection system can continue to mature and become more accurate over time. Additionally, new BAT subtests, measuring other relevant characteristics, can be developed, validated, and added to further improve the way in which candidates are selected for Air Force pilot training.

REFERENCE

Imhoff, D. L. & Levine, J. M. Perceptual-motor and cognitive task battery for pilot selection. AFHRL TR-80-27, AD-A094 317. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, Jan 1981.

SELECTION OF SKILLED MAINTENANCE EMPLOYEES IN THE U.S. POSTAL SERVICE

Henry A. Mueller
Senior Psychologist
U.S. Postal Service
Washington, D.C.

Abstract

The Maintenance Selection System is a comprehensive computer based evaluation system designed to improve the assessment of applicant abilities and knowledge in key qualification areas. The skilled maintenance craft positions filled by this system are in the areas of mail processing equipment maintenance, in addition to building and building equipment maintenance.

The Maintenance Selection System provides procedures, instruments, and techniques to improve the efficiency and quality of maintenance selection and promotion activities in order to ensure that the best qualified candidates are selected for placement into vacant positions. This system integrates valid and reliable independent measures into a comprehensive selection rating of an applicant's total qualifications. Moreover, system generated information is used by maintenance managers to diagnose deficiencies and to provide training suggestions for individual repairmen and technicians in these technical areas.

There are five major elements in the Maintenance Selection System: Qualification Standards, Written Examinations, Review Panel Evaluations, Supervisor Evaluations, and the Scoring and Reporting Program. These elements and the supporting structure and procedures are described.

Introduction

The mechanization and automation of mail processing activities in the Postal Service is increasing at an astounding rate. We now have equipment which can sort more than 32,000 pieces of mail per hour. Building equipment in our facilities is also being upgraded. Complex electronic equipment is used to monitor and control essential environmental systems. The proper, cost effective preventive and corrective maintenance of this intricate equipment is crucial to the operations involved in moving the mail and, indeed, to the continued survival of the Postal Service.

The introduction of this complex electronic equipment has substantially changed the necessary qualifications of those who monitor, maintain and repair this equipment. A number of performance audits which have been completed by both Postal Operations and Inspection Service personnel have indicated that the competencies of the maintenance support staff are less than optimal to meet our expanding maintenance requirements, both in terms of current equipment maintenance activities for which they are given job training, and basic knowledge, skill and ability (KSA) requirements which they should bring with them to the job.

Whenever possible, skilled maintenance craft positions[1] are filled by promotion of qualified applicants from within the maintenance department. If our maintenance operations are to function effectively, we must have a system to determine what the minimum qualifications for a position are; we cannot simply select the best applicants from all those available in the overall labor market.

For most maintenance craft promotions where there are qualified applicants, the best qualified applicant is selected. Qualifications considered include the ability to perform the job, merit, experience, knowledge, and physical ability. However , if there is no appreciable difference in the qualifications of the best of the qualified applicants seniority is the determining factor. If we are to obtain maximum utility we must differentiate between applicant qualifications to the greatest extent possible. This requires that we obtain a comprehensive picture of an applicant's total qualifications for selection purposes.

Hiring unqualified people is costly in terms of both money and time. Productivity can drop. Morale problems can arise throughout the organization. An unacceptably high amount of training may be necessary to develop the basic skills that the new employee was erroneously assumed to bring to the new job. Maintenance managers tell us that in some cases no amount of training can enable the person to achieve even a minimally satisfactory level of performance.

---

[1] In the Postal Service a position refers to the duties and responsibilities performed by one or more individuals doing the same tasks.

The purpose of the maintenance selection development project was to develop the procedures, instruments, and techniques required to select and promote individuals who are qualified and capable of maintaining and learning to maintain our expanding inventory of technologically advanced equipment. In an effort to combine measures covering an applicant's total qualifications into a single selection rating, it was first necessary to identify valid and reliable selection methods which have been found to be predictive of future job performance for similar maintenance positions. A review of the literature looked at previous studies of written examinations as well as alternate selection procedures.

The usefulness of standardized tests for personnel selection is supported by Ghiselli (1966, 1973) in his reviews of published tests. In his 1973 article he states, "As would be expected from the characteristics of the jobs which constitute the trades and crafts, tests of intellectual abilities and of spatial and mechanical abilities have fairly substantial validity for the prediction of trainability." In the same article he goes on to say, "When it comes to measuring aptitude for performing the actual job itself, tests of intellectual abilities, spatial and mechanical abilities, perceptual accuracy, and personality traits are all found to be equally effective, having moderate validity." Ghiselli concludes his article stating, "It will be recalled that single tests are being considered here, and that judiciously selected combinations of tests would have been [sic] higher validity."

Schneider (1976) states "Achievement tests are also used in industry as predictors. Because the nature of an achievement test is to assess the outcome of a process of relatively formal education or formal experience, such tests are useful in predicting immediate job performance." Ghiselli (1966) has demonstrated that an interview rating based on discussion of specific aspects of an individual's previous work and educational history had reasonably high validity, even under very unfavorable circumstances. Schmidt, et. al. (1979) describe an alternative evaluation method which requires job applicants to use structured guidelines in assembling a portfolio of verifiable accomplishments relevant to the job. These accomplishments are then submitted to a panel of job experts who rate the candidate. The rationale for "unassembled examinations" is behavioral consistency: past performance is the best predictor of future performance in similar jobs.

These findings were used to develop a new maintenance selection system which integrates independent selection measures (i.e., written examinations and evaluations of training, education, experience, and job performance) into a comprehensive computer based evaluation system. The developmental process used in the study was designed to ensure that the selection process would be valid, i.e., that it would provide ratings that could appropriately be used to select qualified applicants who would be able to learn and to perform the important tasks in the skilled positions. To assure their usefulness, reliability, and defensibility, the new selection procedures were developed to conform with both professional and federal standards. (APA, 1980; Uniform Guidelines, 1978) This system also enhances uniform application of qualification requirements to ensure the integrity of selection and promotion decisions and to help us meet the Postal Service's merit and equal employment opportunity objectives.

System Development

Job Analysis

A comprehensive occupational study of 54 Postal Service trades and crafts positions was undertaken to establish a data base for the development of a selection program. The initial phase of developing job analysis questionnaires began with a search of existing inventories and previous job analysis studies for similar jobs. Relevant task inventories yielded thousands of elemental tasks which were subsequently refined into 688 more general task statements and grouped into 24 functional job dimensions, called work behaviors. This effort was assisted by the contributions of eight groups of ten initial- and second-level supervisors from diverse postal installations in all five postal regions.

Using self-administered job analysis questionnaires information was collected from approximately 1300 incumbents and the original 80 supervisors. The survey included information about: (1) the tasks and work behaviors performed; (2) KSAs which are required for minimally satisfactory performance; (3) the tools and instruments used; and (4) equipment repaired and types of repair performed.

Measures

Qualification Standards

New qualification standards were developed to describe the minimum KSAs which are required on entry and which are necessary for the satisfactory performance of the duties and responsibilities of each position. They also contain additional KSAs which are not absolutely essential but which contribute to improved job performance by enabling the applicant to perform a wider variety of tasks and/or learn to perform a wider variety of tasks with less training. These are referred to as Desirable Qualification Factors. The Qualification Standards also include examination, physical, training, and licensing requirements.

Written Examinations

Two multiple-choice paper and pencil instruments were developed to measure basic abilities such as mathematics, reading comprehension, and following oral directions. Three multiple-choice job knowledge tests were designed to assess knowledge of such subjects as mechanics, electricity, digital electronics, etc.

Review Panel Evaluation

A Review Panel, established for each position, is composed of these members: (1) the senior installation Maintenance Manager or designee; (2) a supervisor or manager who is knowledgeable in the functional area of the position, and (3) an Employee and Labor Relations representative. The Review Panel evaluates written and oral documentation concerning the candidates education, training, and experience accomplishments pertaining to required and desirable KSAs. Performance Level Scales with behavorial examples, or benchmarks anchor four performance levels on the nine-point scale for each KSA.

Supervisor Evaluation

Nine-point Performance Level Scales with behavorial examples, or benchmarks, are used by the immediate supervisors to describe the applicant's level of qualification on 19 common KSAs.

Major System Elements

Qualification Standards

Based on the job analysis information, new qualification standards were developed for each of the skilled maintenance positions. These qualification standards describe the minimum qualifications as well as additional KSAs which are not absolutely essential but which contribute to improved job performance.

Written Examinations

With the assistance of well qualified incumbents and supervisors with previous maintenance experience, five new written examinations were developed to measure the KSAs required for each of the skilled maintenance positions. An analysis indicated that a majority of the jobs required basic abilities such as math, reading comprehension, and mechanical comprehension. Specific job knowledge is also necessary for satisfactory performance in all of these positions.

Each of the examinations requires approximately three and one-half hours to administer. Once an examination is completed, the KSA scores obtained for all current Postal Service employees are retained in a computer file for use in selection to other positions which require the same examinations. Only the examination questions which cover the required or desirable KSAs for a position are used in the scoring algorithm for that position register.

Except under special circumstances, applicants may not retake an examination for a period of one year. When a Postal employee retakes an examination, the highest score ever achieved on each KSA, on that examination, is retained by the system. There is no penalty for retaking an examination even if a lower score is achieved.

Review Panel Evaluation

A Review Panel is established for each position to evaluate the qualifications of all current Postal Service employees for each position for which they apply. Written applications, supplemental applications and interviews are used to assess the applicant's past training, education and experiences and the effectiveness of the applicant in those situations. The purpose of the assessment is to determine whether the applicant actually possesses the required and desirable KSAs and to what degree.

After reviewing the accumulated evidence each panel member independently assigns a rating to each required or desirable KSA, on the qualification standard, that is evaluated by the Review Panel. The Panel members then

discuss their ratings and assign a composite rating which is the consensus of the Review Panel for each job element. Ratings are recorded on the optically scannable Review Panel/Supervisor Evaluation Form.

Supervisor Evaluation

Immediate supervisors have a major role in deciding who to promote or transfer within the maintenance craft. The supervisor of each maintenance employee who applies for a promotion or transfer evaluates and rates the subordinate's level of qualification on 19 common KSAs which are required for a majority of the positions. Evaluations are made on the basis of the applicant's job performance during the previous twelve months. Performance Level Scales are provided to the supervisors to standardize the evaluation process. If the supervisor cannot rate the subordinate on a particular KSA, due to insufficient information, the supervisor rates the KSA "Not Evaluated" for that individual. This rating does not penalize the applicant because both KSA and register scores are the weighted average of the available information. Ratings are recorded on an optically scannable Review Panel/Supervisor Evaluation Form.

The next-higher level supervisor or manager reviews the ratings and discusses any differences of opinion with the originating supervisor. When concurrence is reached and the forms have been verified they are forwarded to the National Test Administration Center for processing.

A maintenance craft applicant who requests to be re-evaluated by the Review Panel or to retake an examination is also required to have a new supervisor evaluation. This ensures that the most recent information about an employee's work performance is used for the promotion decision.

Scoring and Reporting Program

Application Procedures

Interested Postal Service employees are required to complete a Personal Data/Job Selection PS Form 2518-A (i.e., application) for all positions applied for. All applications for an installation are assembled and submitted to the National Test Administration Center (NTAC) for review. NTAC optically scans the applications and an automated warehouse system returns to the office a complete set of examination materials and a listing of all applicants, indicating those who have examination or evaluation data on file. In addition to examination materials, pre-printed scannable rating forms for the Review Panel and supervisor evaluations are sent to the office. Completed examinations and evaluations are sent back to NTAC for scoring.

When the PS Form 2518-As are received at NTAC a permanent record is created for each employee. Identifying information on this record allows the system to print each applicant's Social Security Number code and other background information on the pre-printed evaluation forms. Evaluation ratings on the forms are automatically matched with the employee's record when the forms are returned to the NTAC and electronically scanned. The employee's record identifies when complete information is received for each position register.

Applicant Results

After the scoring of all materials is completed, a Numeric Eligibility Register (NER) is provided for each position. The NER is essentially a sheet register, or listing of all applicants and their respective register scores. The register score is the arithmetic mean of the KSA composite scores required for the register (i.e., both proficiency and desirable KSAs). The KSA composite scores are the weighted average of the examination score and/or evaluation rating(s) for the KSA. In the KSA composite score the examination score is worth twice the weight of any single evaluation rating. Evaluation ratings on the same KSA are equally weighted.

Every proficiency requirement (i.e., required KSA) for a register has a specific qualifying score which must be achieved before the applicant is eligible for that particular register. All of the applicant's KSA composite scores must equal or exceed the respective qualifying scores in order for the applicant to meet the minimum qualifications and be considered eligible for selection or promotion.

Applicants who do not meet the minimum qualifications are reported as ineligible and are provided with a listing of the KSAs on which they are deficient. In addition, all Postal Service employees, regardless of eligibility status, are provided with a listing of the KSAs on which they scored below the 50th percentile for the incumbent population in the position.

Promotion Eligibility Registers (PERs)

When a maintenance position vacancy becomes available, it is filled first from a PER of current maintenance craft eligibles. PERs are created based on information from written examinations, supervisor evaluations and Review Panel evaluations. The NER score is used to rank order maintenance craft candidates for promotion on the PERs. Candidates who desire to improve their standing on a register may recompete when they have demonstrated the acquisition of new or additional training, education, or experience pertinent to the qualifications for the position.

Inservice Selection Registers

If the PER for a maintenance position is depleted, an office will normally select a Postal Service employee from outside of the maintenance craft. Inservice applicants outside of the maintenance craft are placed on selection registers in the order of their NER scores which are based on their written test scores and Review Panel evaluation ratings. A local Selection Committee (separate from the Review Panel), composed of an Employee Relations representative, and a technical maintenance supervisor in the functional area of the position, is responsible for ensuring that all qualifications, not measured by the written tests or Review Panel, are satisfied before selection or promotion. This is normally accomplished by interviewing previous supervisors, and other knowledgeable individuals to determine whether any negative information exists which would prevent the applicant from performing the requirements of the position satisfactorily.

Entrance Hiring Registers

When no current Postal Service employees are eligible for promotion or reassignment, positions are filled by entrance applicants. Applicants from outside of the Postal Service (i.e., entrance applicants) are ranked on the basis of their written test scores, plus veterans points. Using the rule of three, qualified applicants are considered on an "as-needed" basis in relation to their position on the hiring register. At the time of hiring consideration, applicants must meet the total qualifications for the position. This is determined by a Selection Committee which reviews and evaluates a written and supplemental application for the individual as well as any additional information which they can obtain from applicant interviews and information from previous employers.

Promotion Counseling

The system also provides information which can be useful for promotion counseling. A general comparison can be made between an applicant's test performance and that of the incumbent population. Based on this information, informal meetings with supervisors or training personnel to discuss applicant deficiencies are encouraged. There are many mutual benefits to be obtained by informing applicants on how they can improve and by encouraging them to obtain appropriate training and experience to increase their promotion potential.

Summary

There have been problems selecting qualified maintenance personnel who are capable of maintaining and learning to maintain our expanding inventory of technologically advanced mail processing and building-related equipment. In a coordinated and cooperative effort the Office of Human Resources and Office of Maintenance Management in the Postal Service have developed the Maintenance Selection System in order to ensure that we can select and promote qualified maintenance personnel.

An effective and fair selection system requires accurate identification of current selection requirements, reliable, objective, and valid measurement instruments, and procedures and a timely method of distributing qualification ratings. It is believed that the Maintenance Selection System, using an integrated systems approach and constant surveillance, meets this need for the Postal Service. Current research efforts will monitor system performance and provide a comprehensive understanding of the interrelationships between the measurement methods within the system. Future refinements may enable us to revise and streamline the system and, at the same time, increase the validity and utility in the promotion process.

In conclusion, Ghiselli, Schneider, Schmidt and others have all proposed independent methods for measuring applicant qualifications. The Maintenance Selection System integrates written examinations, Review Panel Evaluations and supervisor evaluations into a composite rating which presents a comprehensive evauation of an applicants total qualifications for promotion to skilled maintenance craft positions.

REFERENCES

Division of Industrial-Organizational Psychology, American Psychological Association. Principles for the validation and use of personnel selection procedures. Washington, D.C.: Author, 1980.

Ghiselli, E. E. The validity of occupational aptitude tests. New York: John Wiley & Sons, 1966.

Ghiselli, E. E. The validity of aptitude tests in personnel selection. Personnel Psychology, 1973, 26, 461-477.

Schmidt, F. L., Caplan, J. R., Bemis, S. E., Decuir, R., Dunn, L., & Antone, L. The behavioral consistency method of unassembled examining. Washington: U.S. Office of Personnel Management, 1979.

Schneider, B., Staffing Organizations (Santa Monica, California: Goodyear, 1976).

Uniform Guidelines on Employee Selection Procedures (1978). Federal Register, 1978, 43, No. 166, 38290-38309.

THE EFFECTS OF THE FLIGHT SCREENING PROGRAM ON ATTRITION IN UNDERGRADUATE PILOT TRAINING

By

JOHN C. QUEBE
Air Force Human Resources Laboratory

INTRODUCTION

Air Force commissioned officer and Officer Training School (OTS) pilot training candidates who do not have a Private Pilot's License are required to complete a Flight Screening Program (FSP). FSP is a 14-hour flying program in the T-41 (Cessna 172). All student sorties are graded by the Instructor Pilot (IP). Students performing especially poorly may be eliminated before completion of the program. After 12 flying hours, students are administered Final Evaluation Flight covering the basic flying skills taught. Students failing this evaluation may repeat it one time. Students achieving a satisfactory Final Evaluation Flight grade proceed to Undergraduate Pilot Training (UPT). All other students are eliminated from the pilot training program for flying training deficiency (FTD). Thus, the program acts as a screen for entry to UPT.

In 1980, an experimental evaluation of the FSP was begun with the aim of determining the effects of the program on UPT and particularly its effectiveness in reducing attrition in UPT below the level which would be expected without FSP. The research plan used to achieve this aim may be regarded as essentially addressing four questions:

1. Does the FSP have any effect on UPT attrition rates?
2. If FSP does affect attrition rates is the effect from screening, training or both?
3. Does the FSP confer a flight training and/or experience benefit?
4. If there is a training/experience effect, would a longer program of FSP flying significantly increase the training/experience benefit?

APPROACH

SUBJECTS

Different FSP treatments were given to different groups of pilot candidates who were then follwed through UPT. Five groups were defined as follows:

Group I (No FSP). Consisted of 123 entrants who would normally have been required to complete the FSP but were allowed to enter UPT directly. These cases, therefore, were unscreened and untrained.

Group II (Extended FSP). Consisted of 57 entrants who were given an extended FSP of 20 hours instead of the normal 14 hours, although screening was still applied at the 14-hour point. These cases, therefore, had 6 extra hours of FSP training and experience.

Group III (Normal FSP). Consisted of 514 students who passed through the normal 14-hour FSP and were screened at the 14-hour point. These cases, therefore, had normal FSP and were screened. This group may be regarded as a control group.

Group IV (Unscreened). Consisted of 266 students who were given 14 hours at FSP but, regardless of performance, were sent on to UPT. In effect, these subjects were not screened.

Group V (FSP Failures). Was defined as a sub-group of Group IV and consisted of 34 of the 266 (13%) who were judged by the FSP IP's to be FSP failures for FTD reasons. Nevertheless, they were allowed to proceed to UPT. The members of this group, therefore, had received 14 hours flying experience but were considered to be unsuitable for UPT.

PERFORMANCE CRITERIA

The FSP treatment effects were evaluated for impact on the following criteria:

1. Pass/fail for FTD reasons at the end of UPT.

2. Pass/fail for all reasons at the end of UPT.

RESULTS

Comparisons were made between different groups to provide the answers to the four primary questions posed in the research plan. Results are based primarily on chi-square analyses. Final UPT training outcomes for each of the five groups are given in Table 1.

TABLE 1. UPT Outcome Data for FSP Groups

	FSP Experimental Group									
	I		II		III		IV		V	
	N	%	N	%	N	%	N	%	N	%
Final UPT Outcome										
FTD	34	28	2	4	78	15	42	16	13	38
Non-FTD Eliminees	20	16	5	8	56	11	39	15	11	32
All Eliminees	54	44	7	12	134	26	81	30	24	71
Graduates	69	56	50	88	380	74	185	70	10	29
Total N	123		57		514		266		34	

OVERALL EFFECTS OF FSP: DOES FSP ACHIEVE ANYTHING?

If the FSP achieves nothing in terms of reduced UPT attrition, it is not acting as a screening device, and is conferring no training/experience benefit.

To determine whether FSP has any affect on UPT attrition, the UPT outcomes of trainees who had entered UPT without previous FSP experience or screening (Group I) were contrasted with those of trainees who had taken the normal 14-hour FSP (Group III). The results of these comparisons are given by phase of training. Table 2 shows that final UPT attrition for all reasons was significantly lower in the group which had been through FSP (26%) than in the group which had not (44%) ($p \leq .001$). Attrition for FTD reasons was also significantly lower (17% vs 33%; $p \leq .001$).

Conclusions. FSP has a significant effect on UPT attrition rates. Students who had been through FSP had lower attrition rates for both FTD and all reasons combined.

TABLE 2. OVERALL (SCREENING AND TRAINING) EFFECT (GROUPS I & III) UPT FINAL OUTCOME

	OVERALL ATTRITION			FTD ATTRITION		
	PASS	FAIL(ALL)	TOTAL	PASS	FAIL	TOTAL
GROUP III (FSP)	380	134	514	380	78	458
(%)	(74)	(26)	(100)	(83)	(17)	(100)
GROUP I (NO FSP)	69	54	123	69	34	103
(%)	(56)	(44)	(100)	(67)	(33)	(100)
Total	449	188	637	349	112	561
	(70)	(30)	(100)	(80	(20)	(100)
$X^2 = 14.33$, df = 1, $p \leq .001$				$X^2 = 12.4$, df = 1, $p \leq .001$		

SCREENING EFFECT OF FSP: DOES FSP SCREEN EFFECTIVELY FOR UPT?

If the FSP screens out probable UPT eliminees, by eliminating them for FTD in FSP, it is fulfilling its primary purpose. Whether the FSP screens effectively for UPT may be approached in two different ways. The first question which may be asked is: "Can individuals who are likely to fail in UPT be identified at FSP?" The second question is: "Does the screening which takes place at FSP significantly reduce attrition rates in UPT?" The first question is concerned primarily with the validity of the FSP as a method of identifying potential UPT failures. The second question is more complex in that the answer depends on organizational factors such as the cut-off standards applied in FSP and the consequent rejection ratios.

Identification of Potential UPT Failures During FSP

To determine whether potential UPT failures could be identified after 14 hours flying at FSP, it was necessary to compare attrition rates in UPT between the 34 FSP 'failures' admitted to training (Group V) and the students who went through FSP alongside them, their contemporaries who were the FSP graduate element of the complete unscreened group (Group IV). (Contempories of Group V were examined to minimize the effects of any changes in FSP and UPT over time.) Through these comparisions, it was found that the overall

attrition rate of the Group V "FSP failure" (71%) was significantly ($P \leq .001$) higher than the overall attrition rate of the Groups IV "FSP Graduates" (25%). Similar results were found with respect to FTD elimination rates in UPT (see Table 3).

Conclusions. These results indicate that some high UPT failure risks can be identified at FSP with a good degree of accuracy, (only 29% of those identified as FSP 'failures' finally graduated from UPT).

The Effects of FSP Screening on UPT Attrition. Comparison of Screened and Unscreened FSP Groups.

To determine whether FSP screening overall had a significant effect on UPT attrition, the UPT outcomes of cases in Group III, who had been through the normal 14-hour FSP and had been screened, were compared with the UPT results of Group IV, who had been through FSP but had not been screened; the latter group contained the 34 'FSP failures' identified as Group V. These analyses revealed no significant differences in UPT results between the screened and unscreened groups. For overall attrition, the rates were 26% for the screened group and 30% for the unscreened group (Table 3). For the FTD attrition criterion, the rates were 17% for the screened group and 19% for the unscreened group (Table 3).

TABLE 3. SCREENING EFFECT (GROUP III vs IV)
UPT FINAL OUTCOME

	OVERALL ATTRITION			FTD ATTRITION		
	PASS	FAIL(ALL)	TOTAL	PASS	FAIL(ALL)	TOTAL
GROUP III						
GROUP III (SCREENED)	380	134	514	380	78	458
(%)	(74)	(26)	(100)	(83)	(17)	(100)
GROUP IV (NOT SCREENED)						
(%)	185	81	266	185	42	227
	(70)	(30)	(100)	(81)	(19)	(100)
TOTAL	565	215	780	565	120	685
(%)	(72)	(28)	(100)	(82)	(18)	(100)
$X^2 = 1.47$, df = 1, P: NS				$X^2 = .14$, df = 1, P: NS		

Conclusions. The implication of these findings is that the 14 hour FSP, with screening at the twelfth lesson did not achieve a significant screening effect for entry to UPT. However, before finalizing such a conclusion two aspects of the research should be noted.

First, the difference in UPT attrition between individuals identified in FSP as high UPT risks and those judged to be better risks has shown that the FSP examiners could discriminate with a resonable degree of accuracy.

Second, while these analyses were unable to show that FSP had a significant screening effect on UPT attrition, there are indications that performance in FSP is related to performance in UPT.

TRAINING/EXPERIENCE EFFECTS OF FSP. DOES THE FSP GIVE A TRAINING/EXPERIENCE BENEFIT IN UPT?

The analyses conducted to identify a screening effect of FSP were not conclusive. The next step was to examine whether FSP provides training/experience which lowers UPT attrition. Possible training and/or experience benefits of the normal 14-hours FSP were examined by comparisons between the UPT results of the group which was not required to attend FSP (Group I) and those of the group that went through the FSP but was not screened (Group IV). Higher UPT success rates for Group IV would be attributable to the training and experience received by this group in the FSP. Overall attrition was significantly ($p \leq .05$) higher in the no-FSP group than in the FSP-experienced group (44% vs 30%; Table 4). Also, attrition for FTD reasons was significantly ($p \leq .01$) higher in the no-FSP group (33% vs 19%; Table 4).

Conclusions. The FSP confers a significant training/experience benefit reflected in the UPT attrition rates. Attrition rates are lower among students who have passed through FSP (even though no screening was applied) than among those who have not been to FSP.

TABLE 4. TRAINING/EXPERIENCE EFFECT (GROUPS I vs IV)
UPT FINAL OUTCOME

	OVERALL ATTRITION			FTD ATTRITION		
	PASS	FAIL(ALL)	TOTAL	PASS	FAIL(ALL)	TOTAL
GROUP III						
GROUP III (TRAINED)	185	81	266	185	42	227
(%)	(70)	(30)	(100)	(81)	(19)	(100)
GROUP IV (NOT TRAINED)						
(%)	69	54	123	69	34	103
	(56)	(44)	(100)	(67)	(33)	(100)
TOTAL	254	135	389	254	76	330
(%)	(65)	(35)	(100)	(77)	(23)	(100)
$X^2 = 6.14$, df = 1, $p \leq .05$				$X^2 = 7.61$, df = 1, $p \leq .01$		

TRAINING/EXPERIENCE EFFECTS OF FSP. WOULD A LONGER FSP GIVE GREATER BENEFIT IN UPT?

If an effect of FSP on UPT attrition is due to training/experience, would a longer FSP course provide greater benefit? This possibility was evaluated next. One group of students (Group II) was given 6 hours extra flying experience at the FSP. Those who had not reached a satisfactory standard by the Final Evaluation Flight were screened out at that point, but the remainder entered UPT with a total of 20 hours of FSP flying experience instead of the normal 14 hours. Differences in UPT attrition rates favorable to this group in comparison to the group which had received the normal 14-hour FSP with screening at the 12-hour point, would be attributable to the extra 6 hours flying experience in the FSP. At the end of UPT, overall attrition was

significantly lower in the 20 hour FSP group (12%) than in the normal-FSP group (26%); $p \leq .05$; Table 5). The difference in FTD attrition was also significant ($p \leq .05$; 4% and 17%, respectively; Table 5).

TABLE 5. EFFECT OF 6 HOURS EXTRA FSP TRAINING (GROUPS II vs III) UPT FINAL OUTCOME

	OVERALL ATTRITION			FTD ATTRITION		
	PASS	FAIL(ALL)	TOTAL	PASS	FAIL(ALL)	TOTAL
GROUP II (20 HR)	50	7	57	50	2	52
(%)	(88)	(12)	(100)	(96)	(4)	(100)
GROUP III (14 HR)						
(%)	380	134	514	380	78	458
	(74)	(26)	(100)	(83)	(17)	(100)
TOTAL	430	141	571	430	80	510
(%)	(75)	(25)	(100)	(84)	(16)	(100)
	$X^2 = 4.53$, df = 1, $p \leq .05$			$X^2 = 5.18$, df = 1, $p \leq .05$		

SUMMARY OF RESULTS FOR THE EFFECTS OF EXPERIMENTAL FSP TREATMENTS ON UPT ATTRITION

(1) The analyses showed that the current 14-hour FSP had significant effects on attrition in UPT. Attrition rates were lower in the group which had undergone the FSP (Group III) than in the group which had not (Group I).

(2) High UPT-attrition risks could be identified in FSP. However, with the data available for analysis, no difference in attrition rates in UPT was apparent between the group which had been screened in the FSP (Group III) and the group which had not (Group IV).

(3) There was clear evidence that the FSP conferred a significant flight training and experience benefit. Attrition rates in UPT were lower among pilots who had taken the 14-hour program without being screened than among students who had been allowed to enter UPT directly.

(4) Extension of the FSP to 20 hours of flying gave an additional training benefit. Attrition rates in UPT were significantly lower in a group which had received the extended FSP than in the group which had undergone the 14-hour FSP.

WHAT PERFORMANCE DOES A PERFORMANCE TEST TEST?

T. M. Ansbro
Staff, Chief of Naval Education and Training
Naval Air Station, Pensacola, Florida

Performance testing addresses performance of work by job incumbents and work planned for prospective incumbents, worker requirements for standing or certification in occupational fields, and/or skills profiles for personnel and manpower models. Our attempts to measure or otherwise evaluate these performances in productivity assessment, performance evaluation, or other process/product measurement subsystems don't always measure appropriately, completely, or expertly.

An American executive recently said, "Nobody in management feels comfortable having to deal forthrightly with embarrassingly true, painfully accurate, detailed, and technically verifiable performance evaluation until the stakes are so high that there is no alternative." He then cited the highly successful landings of the Space Shuttles, every one performed before the world public, and frequently a first-time effort by each pilot. Consider the "consequence of inadequate performance" (favorite job/task questionnaire item): unacceptable risk or loss of life, astronomically expensive equipment, and almost irreversible damage to US technology prestige. Performance requirement: near-perfection, passing score 100% (minus perhaps 2% margin for error), failure inconceivable, and job-performance environment "go, no-go".

The majority of the world of work enjoys a job-performance environment devoid of most of such harsh realities and dangers, more than a bit less demanding, and certainly more relaxed. When consequence of error is less than stringent, we may relax a bit too far, lose some perspective in establishing and applying performance standards to language of evaluation, or lose a skirmish in the battle of bias and ignorance versus informed objectivity. Consider these examples of executive judgement:

US Federal Civil Service --- "His performance is outstanding in every respect" --- rater awarded to 50% of his staff each year for several years.

US Navy printed personnel rating form (50th percentile) --- "Typical Outstanding 4.0 Chief Petty Officer". Possibly an inspiration for this comment from a Commanding Officer of an activity --- "I don't like to think of these occupational standards (OCCSTDS) as 'minimum' requirements of the rate. What we want in the Navy is performance ABOVE the minimum. I want everybody in this outfit to be 'Outstanding'. 'Satisfactory' is 'High Lousy'. Everybody knows that."

Last year's College Board's specification of needed learning outcomes in English Literature cited "ability to read critically, read literary text analytically, read with understanding a range of literature rich in quality and representative of different

BEST COPY AVAILABLE

literary forms and cultures" --- specifically, "interest in and a sense of inquiry about written words, ability to respond intuitively and imaginatively to literature."

A conference critic of Instructional Systems Development (ISD) cited "ISD Overkill" --- an ISD analysis of "skills" required to drive a car turned up 4,000 "discrete behaviors." The presenter opined that there are "only a handful of general competencies" needed to master driving a car "successfully;" therefore, "these are the ones to define and teach." Precisely how to get to that handful of competencies was not presented.

More examples of this kind might tempt a cynical observer to wonder how we ever evaluate with reasonable accuracy what we do or assess the product of our endeavors; a concerned one might concede the mess and ponder how to cull through, examine our world of work, and come up with an appropriate matrix of performance measurement. He might determine that we need to straighten out an evaluation vocabulary of such terms as (above) "generic learning outcomes, general competencies, quality, typical, outstanding" with such disparate quantities as several thousand versus "a handful," or percentile ratings not clearly supported by concrete component scaling. Such terms must beg for clarification if not uniformity.

What are the objectives of work evaluation in the first place? Design a job, describe it, hire/train a prospective incumbent, transfer/promote/retrain an incumbent, assemble an occupational field? Construct a ship/squadron/batallion manning document? Certify or license members of an occupational/career field? Confer credentials? Develop curricula/training programs? Whatever the objective, there should be a definitive test to establish that it has been or can be met. Methodologies and a variety of instruments abound, within fields and subdivisions of fields, accompanied by job-knowledge data, also, categories: current US armed forces job classfications (MOS, AFSC, Rating, NEC) number some 4200, with additions pending. We enjoy an embarrassment of riches, but not a uniform system to put all these riches together in a most meaningful way.

And that suggests that in terms of job performance evaluation there should be a test for each identifiable element of work that we can logically assemble within the categories with which we are dealing: occupational field, job, duty, task, skill, certification/paygrade requirement, and circle of standardization. These ingredients should support any testing program with objectives to determine or establish:

Performance requirements for a prospective jobholder to meet (job entry), satisfactory on-job task/skill performance;

Skill mastery levels, task/skill ranking criteria;

Job/skill certification/licensing criteria.

Needed are coded occupational-field and job-knowledge data to support the above; also, work-behavior inventories for projected or incoming technologies, systems, equipment (and task/skill performance deltas between them and contemporary counterparts).

BEST COPY AVAILABLE

Choosing the ingredients to provide a performance test for such objectives requires having these ingredients at our disposal: appropriate inventories, fleshed out, current, and with requisite analysis mechanisms to provide appropriate data outputs --- a job for the computer. Fortunately, many of the requisite mechanisms are either in development or in use. Logistics Support Analysis (LSA) is a massive and ambitious systems/equipment acquisition and occupational data base already in at least incipient function under a DOD mandate. HARDMAN, which should adjunct to LSA, is also under way in the Navy, and an adaptation (MANPRINT) is in the works in the Army.

What work behaviors are germane to all these objectives? Of what use are specific skill elements as determinants of task (eventually, job) performance capability? Is there a clear line from skill demonstration to task performance; an "audit trail"? Can task performances be amalgamated to indicate job performance, job performance scientifically grouped to determine competency requirements for an occupational field? Information recall (of fact, rule, or principle), frequently labelled "remembering", is a conventionally required test-item mechanism. How often is it really the work behavior most appropriate for testing? Calling Morse Code out of memory is an enabling mechanism (skill) for sending/receiving/encoding/decoding messages (tasks), which also employ other computational and manipulative skills; it may be buried deeply among the suite of corporate skills that in the aggregate demonstrate worker capability to perform these tasks. Potential to predict performance capability would be relative to the skill's prominence among the aggregate; and only where skill prominence is appropriately reflected in a test of performance would its contribution to capability predictability be valid.

Figure 1 is an attempt at representing flow, and in a sense, metamorphosis of basic information ("job knowledge") through skills toward job task performance. Where the evidentiary links between components in the task-internal hierarchy are unbroken, there would be an audit trail establishing (or at least tracing) the connection, and the contribution, of static and processed information to skills performance and on up to job task performance (Ansbro, Hayes, 1981). Without such an audit trail, the mere existence of knowledge items in an inventory of task-supporting references is not sufficient to justify their selection for use in prospective test items to predict or reflect performance; nor is random selection among task-underlying skills. Appropriate reflection of the skill most closely associated with task performance is the keyword.

It may appear satisfactory to test worker speed and accuracy in sending and receiving, since these are the job tasks, and we can deduce that task-component skills have then been inferentially tested. But, for a proper profile of the task and skill performance suite of a classic worker in a model job in a recognized occupational field (for training program development,

position description, or manning document construction), we need performance information on more than obvious terminal tasks performed on a specific job. We may have to test for skills identified as transferable. In the aggregate, it is the skills that are transferable among and across tasks. Task requirements are specific, therefore fixed; skill requirements tend toward the generic, are therefore fluid elements in a performance matrix.

In a rough graphic representation of a single-job task/skill/knowledge inventory, a job may be depicted as in Figure 2. At this point, task subdivisions (Omnibus, Embodied, Unique) in Figure 1 acquire significance. Omnibus tasks contain all component work behaviors (skills, etc) of action-associated tasks of lesser complexity and smaller compass in the inventory; therefore "embody" them. "Unique" tasks have no such direct connections of commonality among at least the majority of their component work behaviors (Davis, Perry, 1980; Ansbro, 1984).. Illustrated is a single job encompassing its duties and tasks, and their interrelations. Omnibus tasks encompass embodied tasks, and in turn, support performance of duties; the whole constitutes a job.

Skills underlie and support many tasks, but assume varied importance or criticality to task performance; they extend through multiple layerings of task performance and transfer not only across tasks but also across jobs and occupational fields. Soldering and welding skills, for instance, provide extensive task support throughout metalworking, electricity/electronics, and fabrication occupational fields. Much the same characteristics attend skills associated with use of general purpose test equipment. Skill layerings (or mastery indexes) within these skills apply variously among task hierarchies; and to depict this increasingly complex environment of task-skill-occupational field interrelationship requires a matrix.

Expand the single-job illustration by replicating it in a single lateral line as far as established or assumptively associated job-task-skill boundaries extend, and we have a picture of an occupational field. Not every component of that field will be equally typical, but a circle of standardization can be established that will encompass what is most typical or representative. Relatively esoteric, extra-high-tech, or rarely performed jobs or tasks (sometimes including skill support) will fall outside the circle, but still within the field. It's a short step from this depiction to laying out a matrix of field/job/duty/task/skill relationships and skill-mastery layerings from which arrays of work-behavior groupings may be assembled to provide performance test action items and evaluative criteria. With adequate occupational data base support, the computer can either be called in to assist in the make-up of the testing, or it can be programmed to do the job itself.

With respect to testing in an occupational field, Navy

custom has been to examine for job-knowledge competence in a field rather than in a specific job within that field (or rating), therefore, generally within the circle of standardization. This exercise, however, becomes more difficult as the fields expand with the constant advance of technology, which, as it becomes more platform/system/equipment-specific, may be narrowing the circle to the extent that it represents less of the field than can still be considered authoritatively representative. Such almost-amoebic reproduction of jobs and tasks can periodically (or at least eventually) require redefinition of the field. But it is the specifics of job and task descriptions that tend to broaden and fractionate the occupational fields, while the common component (and therefore transferable) skills tend to hold the fields in still-recognizable form. However, increasingly expanding application of the computer to large data fields should tell us that it is no longer necessary to hold tightly to established structures when they are expanding or changing to differing configurations.

Establishing job performance requirements for testing or the objectives discussed herein necessitates a massive, up-to-date occupational data base of sufficient descriptive detail and inventory depth to store, analyze, and rank data input, sort for job/duty/task/skill commonality across and among worker classifications, and establish task and skill complexity and componency indexes (Ansbro, 1984). Such a data base and analysis mechanism must function across and among job classifications and be internally blind to occupational specialty insignia. Task and skill ranking should depend upon complexity and componency; commonality and uniqueness should apply across and among classifications, for definition or redefinition of occupational field boundaries. Job/duty/task/skill/knowledge data should be cross-coded within, among, and across occupational fields and training programs, and should be retrievable elements in hardware and systems/equipment acquisition data systems, involving training/learning objectives in courses of instruction and the structure and components of performance testing.

REFERENCES

Ansbro, T. M., The Case For A Common Data Base. Proceedings, Second Annual Air Force Conference on Technology in Training and Education (AFTITE). Sheppard AFB, TX: 1984

Ansbro, T. M. and Hayes, W. A., The Job Task Analysis/Skills and Knowledge Marriage. Proceedings, 23rd Annual Conference of the Military Testing Association. Arlington, VA: 1981

Davis, D. D. and Perry, N. N., Determining Task Commonality in Navy Training. Proceedings, 22nd Annual Conference of the Military Testing Association. Toronto, Ontario, Canada: 1980

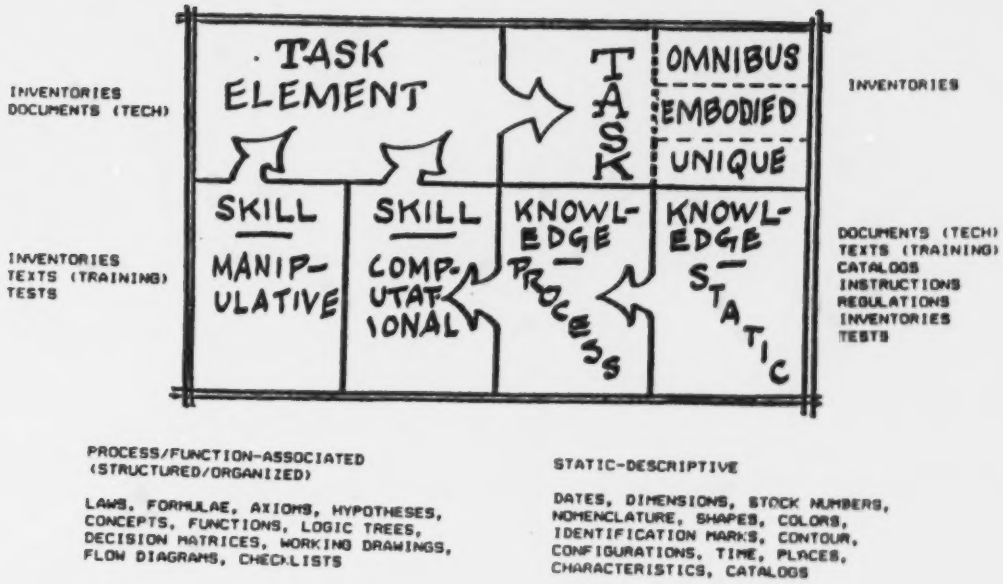


Figure 1.

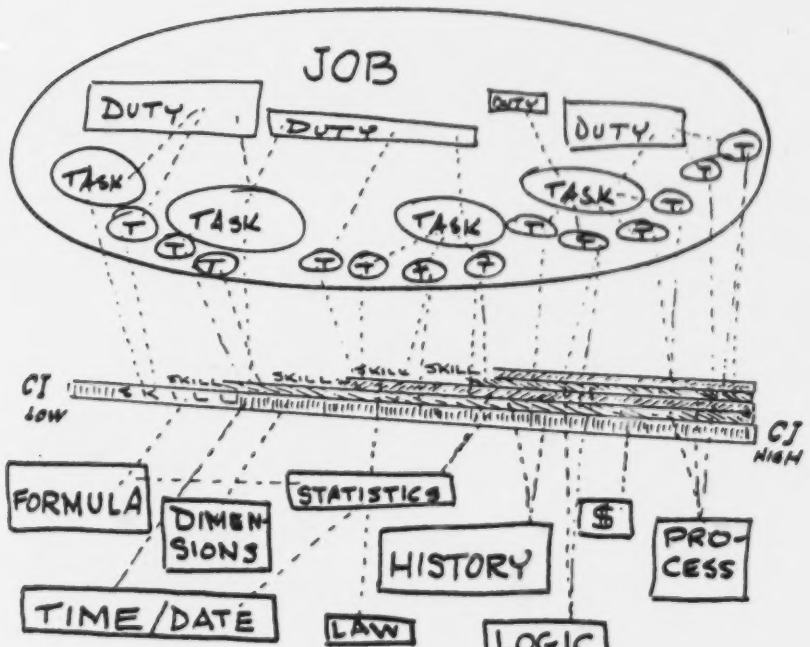


Figure 2.

How Qualitatively Informative are Test Items?: A Dense Item Analysis

William M. Bart
Educational Psychology
University of Minnesota

The concept of the "dense item" was first introduced in a paper presented by W. Bart at the 26th Annual Military Testing Association Conference held in Munich in 1984. This concept is central in a framework directed to the goal of improved diagnostic and prescriptive testing. The dense item framework and its associated methodology of refined item digraph analysis has subsequently and unexpectedly provided a way to evaluate cognitive theories and instructional theories. This paper describes two quantitative indices that can be used to tell the degree to which a test item has the first two qualitative properties of a dense item.

The Dense Item

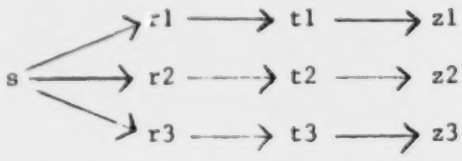
A dense item is any test item that indicates exactly why subjects provide the responses they give and exactly what instructional sequences should be provided the subjects to correct any faulty rules or procedures they may be using. Thus a dense item is an ideal diagnostic and prescriptive instrument and serves as a bridge between the field of learning and cognition and the field of instruction and teaching. A dense item has ten properties and each property has its own quantitative index. The first two properties are response interpretability and response discrimination. These two properties and their indices are featured in this paper.

The Refined Item Digraph

Bart (1984) introduced a methodology termed "diagram analysis" and intended to investigate test items. This methodology has been refined and relabelled as "refined item digraph analysis". The refined item digraph is a graphic way to depict the associations between an item stem (e.g., "2 + 2 = __") and the responses to the item stem (e.g., "4") and the inferences among the responses, the rules that relate the item stem to the responses, and the instructional sequences that relate the rules, whether defective or not, to the correct rule. Psychologically, a rule is a sequence of one or more cognitive operations that permits an individual to generate a response from an item stem and an instructional sequence is a sequence of one or more instructional experiences that permits a student to learn the correct rule from an initial mastery of a defective rule.

A refined item digraph of a test item with three responses, if the item were dense, would use the following notation: (a) the item stem is termed "s" and constitutes a set S; (b) the responses are termed "r1, r2, and r3" and constitute the set R; (c) the rules are termed "t1, t2, and t3" and constitute the set T; and (d) the instructional sequences are termed "z1, z2, and z3" and constitute the set Z. In this case, let us assume that r2 is the correct answer, then t2 is the correct rule and z2 is the identity instructional sequence which, when employed, maintains the knowledge and usage of the correct rule. The refined item digraph of this item if it were a dense item would then be the following:

Figure 1: Refined item digraph of a dense item with three responses.



This refined item digraph indicates the inferences an instructor could make from a consideration of the responses of a student to the item. For example, if a student generated r3 as his response to item stem s, the teacher would know not merely that the response was wrong, but also that r3 results from usage of defective rule t3. The teacher could also infer that instructional sequence z3 should be provided to the student so that he/she can learn t2, the correct rule.

A refined item digraph has only between-set inferences and no within-set inferences interrelating the sets S, R, T, and Z for an item. An item digraph has, however, both between-set inferences and within-set inferences interrelating the sets S, R, T, and Z for an item and being indicated by arrows. A refined item digraph and an item digraph are both digraphs, because they both are arrays of points interconnected by arrows (Harary, Norman, & Cartwright, 1965).

In the explication of the dense item and its properties, it is sufficient to consider only between-set inferences for an item and that is why it is parsimonious to analyze only refined item digraphs and not item digraphs in general for this exposition on the dense item.

The Index of bod

One important concept in digraph theory that is useful in refined item digraph analysis is that of between-set outdegree. The between-set outdegree of point x, bod(x), is the number of arrows emanating from x to other points in other constituent sets in the item digraph. As an example, bod(r1) = 1 for Figure 1, because only one arrow emanates from r1 in set R to points in the other sets S, T, and Z. That one arrow indicates an inference from response r1 to rule t1. This index of bod will be crucial in defining the quantitative indices associated with the first two properties of the dense item.

Two Properties of the Dense Item

The first two properties of a dense item are response interpretability and response discrimination. In the subsequent sections, each property and its quantitative index will be described. In addition, an example will be provided regarding the computation of the index.

Response Interpretability

A test item has response interpretability if each response to the item is interpretible by at least one rule. An index of response interpretability, I1, indicates the degree to which an item has response interpretability.

In order to define $I1_i$, the index of response interpretability of item i, it is necessary to define $I1_{ij}$, the index of response interpretability for response j to item i. $I1_{ij}$ is defined in the following manner:

$I1_{ij}$ = 1 if $bod(r_{ij}) > 0$ and (1)

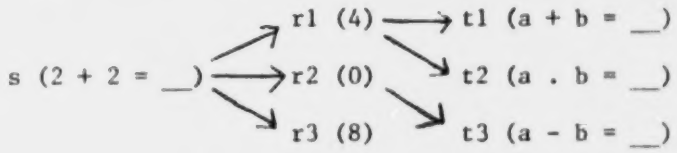
= 0 otherwise.

$I1_i$ is then defined in the following manner:

$$I1_i = \sum_{j=1}^{k} I1_{ij}/k \text{ with k being the number of responses to item i.} \quad (2)$$

To exemplify the computation of these indices, let us consider the following item involving the item stem "2 + 2 = __", the responses "4, 0, and 8", and the rules "a + b = __, a . b = __, and a - b = __." With this item the correct answer "4" is interpreted by the correct rule "a + b = __" and by the incorrect rule "a . b = __", the incorrect answer "0" is interpreted by the incorrect rule "a - b = __", and the incorrect answer "8" is uninterpreted. Figure 2 depicts the refined item digraph for this item.

Figure 2: Refined item digraph of an item with three responses.



The index of response interpretability for this item can be calculated in the following manner. In this item, k = 3, because there are three responses to this item.

$$I1_i = \sum_{j=1}^{k} I1_{ij}/k$$

$$= (I1_{i1} + I1_{i2} + I1_{i3})/3$$

$$= (1 + 1 + 0)/3$$

$$= 2/3$$

$$= .67.$$

The index of response interpetability for this item is .67 which indicates that 67% of the responses to the item are cognitively interpreted. This value is quite high and would very likely be superior to the indices of response interpretability of most other items. Most items have little response interpretability and thus have little diagnostic value.

Response Discrimination

A test item has response discrimination if each response to the item is interpretible by only one rule. The index of response discrimination, $I2_i$, indicates the degree to which an item has response discrimination.

In order to define $I2_i$, the index of response discrimination of item i, it is necessary to define $I2_{ij}$, the index of response discrimination for response j to item i. $I2_{ij}$ is defined in the following manner:

$$I2_{ij} = 1/bod(r_{ij}) \text{ if } bod(r_{ij}) > 0 \text{ and} \qquad (3)$$

$$= 0 \text{ otherwise.}$$

$I2_i$ is then defined in the following manner:

$$I2_i = \sum_{j=1}^{k} I2_{ij}/k \text{ with k being the number of responses to the item.} \quad (4)$$

To exemplify the computation of these indices, let us apply them to the item depicted in Figure 2.

$$I2_i = \sum_{j=1}^{k} I2_{ij}/k$$

$$= (I2_{i1} + I2_{i2} + I2_{i3})/3$$

$$= (1/2 + 1 + 0)/3$$

$$= 1/2$$

$$= .50.$$

The index of response discrimination for this item is .50 which indicates a modest level of response discrimination for this item. If a response is either uninterpreted or is interpreted by more than one rule, then that response does not discriminate well among cognitive rules. As the number of rules which interpret a response increases, so does the degree to which the response discriminates among cognitive rules decrease. That condition is reflected in the formulation of equation (3). The index of response discrimination is an item discrimination index which is concerned with the degree to which an item discriminates among cognitive rules and not ability groups as is the concern of item discrimination indices of conventional psychometric theory.

The Promise of Refined Item Digraph Analysis

Refined item digraph analysis provides a psychometric system which can be used to evaluate the extent to which an item has the dense item properties and thus has diagnostic and prescriptive value. This approach to item analysis involves ten dense item properties and has been developed by W. Bart. Refined item digraph analysis also provides a way to evaluate cognitive theories and instructional theories. Although only two dense item properties are discussed in this paper, it is hoped that these two properties and their quantitative indices can be used to evaluate the diagnostic values of existing items and to suggest areas of item research which could lead to improvements in the diagnostic values of test items.

References

Bart, W. (1984). The dense item: a bridge between learning and instruction. Proceedings of the 26th Annual Conference of the Military Testing Association. (pp. 391-396). Munich, Federal Republic of Germany.

Harary, F., Norman, R., & Cartwright, D. (1965). Structural models: an introduction to the theory of directed graphs. New York: Wiley.

Acknowledgment

The author wishes to express his appreciation to the following parties for their financial assistance and support in this research: (a) Prof. Dr. F. E. Weinert and the Max Planck Institute for Psychological Research; (b) the Fulbright Kommission; and (c) the Wilson/ U. of Minnesota College of Education Alliance.

Test-Item Readability: How the Variables Work

R. Eric Duncan, Captain, USAF
University of Texas at Austin

Introduction

In 1981, Duncan proposed a model of test-item readability which incorporated item, examinee, and environmental characteristics. These characteristics interact to produce the readability level of an item. Duncan (1981), however, lacked a solid theoretical framework from which to relate the semantic and syntactic components of an item to its readability. To correct that deficiency Duncan adapted Kintsch's (1974) propositional approach to text-based readability to multiple choice test items. This paper will briefly describe this adapted approach, describe the most critical syntactic and semantic variables which predict item readability, and present initial results in the attempt to predict item readability.

Semantic Variables and Kintsch's Propositional Theory

The variables described here can be grouped into four distinct areas: semantic variables, syntactic variables, a cognitive load variable, and measures of prior knowledge. The semantic variables include propositional density, operator density, argument density, and propositional level. The variables in this area are directly related to Kintsch's (1974) propositional approach to the description of memory in semantic memory and so a brief explanation of that theory is in order.

Kintsch's (1974) semantic approach to reading and processing textual material focuses on the proposition. Kintsch and Keenan (1973) point out that sentences read from text are not stored verbatim, but rather as propositions. Propositions are word concepts combined to form a logical set of lexical items and contain a relation (usually a verb) and n arguments (nouns, adjectives, pronouns). These propositions are put together in a logical manner, establishing a representation of the text, known as a text base, in memory. A text base is simply "an ordered list of propositions" (Kintsch, 1974, p. 13). To obtain this list of propositions, the text must be semantically analyzed. Propositions, which include relation and argument(s), are then abstracted beginning with the first sentence.

Before describing the propositional construction process, relations and arguments need to be described. A relation is a word concept (not necessarily a word) which describes some action or state of being and normally appears as a verb, adjective, adverb, or noun.

John sleeps	(SLEEP, JOHN)	(1)
Mary bakes a cake	(BAKE, MARY, CAKE)	(2)

In example (1) from Kintsch (1974), SLEEP, a verb, is the relation describing some action that is being performed by the argument, JOHN. Fillmore (1971) established semantic rules for arguments. Arguments must be an agent, experiencer, instrument, object, source or goal and are structured in that order of importance. In example (2), MARY is the agent that BAKE(s) the object CAKE. Propositions can serve as arguments for other propositions, as well. In example (3),

If Mary trusts John,	(TRUST, MARY, JOHN) = a	(3)
She is a fool	(FOOL, MARY) = b	
	(CONDITION: If, a,b) = c	

the proposition "c" has propositions embedded in it as arguments. This function is useful when building text bases, since it is more economical and requires less memory space than recreating new propositions that had been processed earlier.

BEST COPY AVAILABLE

Levels and Operators

Two additional important features in the derivation of a text base are presented in example (3). The first feature is that of the level of propositions. "A proposition is said to be subordinate to another if it contains an argument that also appears in the first proposition" (Kintsch, *et al.*, 1975). Subordination can occur, as is most often the case, immediately after the superordinate proposition in the text base, or can occur much later in the text base, as in the case of propositions being used as arguments for other propositions. In example (3), subordination is indicated by the indentation of propositions "b" and "c". Indentation of propositions is a convention established by Kintsch and was used in creating item protocols.

The second important feature shown in example (3) is that of operators. Operators are mechanisms which require inference on the part of the reader. Simply, operators do not state an explicit relationship among propositions, but rather challenge the reader to obtain an implicit meaning from preceding information. Operators can also be described as taxing memory space and requiring memory searches to find them that and how to manipulate previously encountered propositions. They include: Causality, Contradiction, Part, Time, Location, Condition, Conjunction, and Purpose. The operator would normally appear in the listing of the text base as presented in the "c" proposition. One additional operator (MATCH) has been created that was not included in Kintsch's operators. This operator identifies the cognitive operation of matching the propositions in the item alternatives to those propositions stored in memory.

Empirical Evidence in Support of Kintsch's Theory

Now that the basic components of Kintsch's theory have been described, this section provides the empirical evidence which supports the contention that test items are broken down into logical semantic units (propositions) before storage in memory.

Kintsch and Monk (1972) demonstrated that experimental subjects stored text material in the same manner, regardless of the syntactic complexity of the text. They found that the more syntactically complex paragraphs took longer to read but that there was no significant differences in the number of propositions recalled. Kintsch and Monk suggested that text is not represented syntactically in memory, but rather that it is represented semantically in propositional form. This evidence supports the contention that text is parsed into propositions and is stored in semantic memory.

Kintsch and Keenan (1973) examined the effect on reading time and recall of the number of propositions and the level in text of the propositions. The length of sentences (total number of words) and the number of propositions they contained were covaried. This approach is better known as propositional density, i.e., number of propositions/number of words. The levels of propositions were also varied. Kintsch and Keenan found that, if reading time was unlimited, propositional density significantly affected recall rates. They also demonstrated that superordinate propositions were recalled better than subordinate propositions. This finding was later supported by Kintsch, *et al.* (1975). This result can be more easily explained by referring back to example (3). Proposition "a" is a superordinate proposition while propositions "b" and "c" are subordinate to "a". Kintsch and Keenan, and Kintsch, *et al.* have shown that proposition "a" has greater probability of recall than propositions "b" & "c".

In addition to supporting the results of Kintsch and Monk (1972) and Kintsch and Keenan (1973), Kintsch, *et al.*, (1975) also examined

BEST COPY AVAILABLE

the effects that the number of different word concepts (arguments) would have on recall. Results indicated that the more frequently a word concept (argument) is repeated in the text, the better it is remembered. The authors also showed that, as the number of different arguments increases, reading time increases and recall decreases. Kintsch, *et al.* suggested that the history paragraphs may be easier than the science paragraphs they used because they contain propositions that "are already part of the subjects' general knowledge" (p. 209).

Variables of Interest

Experimental evidence has shown that propositional density (# of propositions/# of words), operator density (# of operators/# of propositions), argument density (# different arguments/# of propositions), and propositional level contribute to reading comprehension. The hypothesized relationship between propositional density (PD) and reading comprehension scores in test items was that as PD increased (more propositions per word), the reading score necessary to understand the item would also increase. This is also true for operator density (OD), argument density (AD), and propositional level (PL).

Syntactic variables from Duncan (1981) [illegible] be included in the estimation of test item readability. Centerembeddedness, a modifying word or phrase located between the subject and predicate of an item, has been shown to make text less comprehensible than present (Lambert and Siegel, 1974). A modifying phrase that precedes the subject of an item is known as a left-branched phrase. A phrase which follows the predicate of an item is known as a right-branched phrase. Schwartz, *et al.*, (1970) showed that left-branched phrases reduced the comprehensibility of text while right-branched phrases had no appreciable effect on comprehensibility. These three variables, then, constitute the syntactic element of test-item readability.

There are two other elements that contribute to the prediction of item readability: (1) measures of prior knowledge, and (2) cognitive load. Kintsch and Vipond (1977) suggest that prior knowledge could enhance a person's ability to extract meaning from text. They indicate that the best method to assess prior knowledge is with a vocabulary test containing words used in the text material being read. This direct method, in the present study, is impossible to apply since all test and subject data had been taken from historical data files. Substitute measures include jargon and uncommon words. The number of uncommon words (UW) is determined by comparing all words in an item to the *Common Word List* compiled by Kincaid, *et al.*, (1980). This list contains the 20,000 most common words used by enlisted Navy personnel with a 9th grade reading ability. After comparison, those words not appearing in the list are compared to the text reference material to determine if they are sufficiently explained. If there is no explanation or definition given, the word(s) are counted as uncommon. The second variable used to assess prior knowledge is jargon. Jargon is based on the judgement of the item writers'. Jargon words, such as CBPO, grade, and MAJCOM are not common to the general public and are specific to Air Force personnel. The jargon variable is expected to enhance readability since the use of jargon is commonly used in the military to communicate frequently complex names or phrases in a succinct manner. The variable used to measure cognitive load is Bloom's Taxonomical Level (Bloom, *et al.*, 1956). This variable indicates the cognitive activity necessary to read and answer a test question and includes the following levels: (1) rote memory, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6)

BEST COPY AVAILABLE

evaluation. Raters will evaluate each item on this variable and indicate the appropriate level.

The variables of interest, then include four semantic variables adapted from Kintsch's propositional approach, three syntactic variables, two measures of prior knowledge, and a variable to assess cognitive load.

Experimental Approach

One hundred multiple-choice items from a commonly used Air Force test were evaluated on the variables of interest by seven trained raters in a counter-balanced fashion. The data was reviewed by the author to insure accuracy. Two thousand examinees were randomly selected and their Armed Services Vocational Aptitude Battery scores were used to determine each examinee's reading grade level (RGL). This was accomplished using regression equations provided by Madden and Tupes (1966). The RGL of the 25th percentile of all examinees answering each item correctly was assigned as the RGL for that item.

The RGL and *p*-value for each item were used as the dependent variables in further analyses. Regression analyses were performed to determine if statistically significant equations to predict test-item readability and item difficulty could be produced. Two regression approaches were attempted - all possible subsets regression and stepwise multiple linear regression. The independent variables were the item variables previously described.

Results and Discussion

Frequency analyses showed diversity among predictor variables with the syntactic variables providing little variability. This occurred because the items evaluated were relatively syntactically simple. Eighty-three percent of the items had no center-embedded phrases and seventy-eight percent had no left-branched phrases. RB did show a bit more variability with a range of 0 to 8 right-branched phrases, a mean of 1.92, and a standard deviation of 1.73. Seventy-nine percent of the items were evaluated at the rote-memory level of Bloom's Taxonomy and ninety-six percent of the items were at or below the application level. Forty-seven percent of the items contained at least one jargon word with one item containing six different jargon words. Since these items were written at a relatively simple level, only sixteen percent of them contained one uncommon word, three percent with two, and one with three uncommon words.

The semantic variables seemed to have more normal distributions with acceptable means and standard deviations. The information shown below provides the means, standard deviations, variances, and numbers of cases. It should be remembered that, except for propositional level, these data are proportions, with a range of 0.00 to 1.00.

Variable	X	S.D.	Variance	N
Propositional Density	.609	.160	.026	100
Operator Density	.501	.100	.010	100
Argument Density	.583	.116	.013	100
Propositional Level	5.430	1.183	1.399	100

These variables provide the most variability for use in correlation and regression of all the independent variables.

The 25 percentile RGL criterion variable had less range and variability than anticipated. This variable had a mean of 8.916, standard deviation of .123, and a range of 8.691 to 9.356. The item *p*-value criterion had a mean of .470 and a standard deviation of .2068, with the most difficult item being .08 and the easiest .93.

BEST COPY AVAILABLE

Evaluation of the correlation matrix is made with a 0.10 level of significance in mind. None of the correlations among the syntactic variables (CE, LB, and RB) were significant. This is most likely due to the lack of variability in CE and LB. The relationships of the syntactic variables and the measures of prior knowledge were also insignificant. However, the correlation of -.19 between jargon (JAR) and uncommon words (UW) was significant at p=.06. If the theoretical relationships between these variables and comprehension are true, the negative relationship indicates that jargon aids in item comprehension, while uncommon words hinder.

The correlations between syntactic and semantic variables tended to support anticipated relationships. The correlation between propositional density (PD) and RB ($r=-.42, p<.001$) indicates that as the number of propositions per word increases the frequency of right-branched phrases decreases. This finding is logical, in that, high PD values usually occurred in items with few words. The propositional level (PL) of an item often shows modification within it. RB also indicates modifying phrases. This a strong relationship should and does exist. Surprisingly, operator density (OD) has a positive correlation with LB ($r=.23, p<.05$). This result indicates that as the number of operators increase so does the number of left-branched phrases. The only explanation the author can provide is that OD, like LB may contribute to less comprehension as its frequency increases.

The correlations among the semantic variables indicates that argument density (AD) has stronger relationships with the other semantic variables than does PD. This may be a function of the lack of linearity in PD. The OD variable relates negatively to AD which reinforces the previous observation about OD ($r=-.51, p<.001$).

The regression analyses were disappointing in that the R^2 values accounted for little variance. In the stepwise multiple linear regression analysis with the 25th percentile RGL as the criterion, only three variables contributed to a meaningful change in the R^2 - RB, LB, and Bloom's Taxonomy. The regression equation with these variables produced a multiple R of .231 and an R^2 of .053 with an F of 1.75 (nonsignificant). When stepwise regression was used to predict item difficulty, different independent variables contributed to explained variance. These variables, jargon, PD, and OD, produced a multiple R of .232 and an R^2 of .054, with an F of 1.77. The all possible regressions analysis, using R^2 as the selection criterion, produced almost identical results.

While neither of the analyses with either criterion variable produced a significant R^2, they do point out some possible trends. One of the more interesting trends is that syntactic variables tend to predict RGL better than semantic variables and prior knowledge, while semantic variables and jargon tend to predict item *p*-values better than the other variables. This author believes that measured RGL (or estimated in the case of the criterion variable used in this research) may reveal how an item is put together and not what it is asking. For example, a person with a higher RGL may be better able to grapple with complex syntax and decide what type of information is being sought (rote, application, etc.) more easily than an individual with a lower RGL. This then allows the better reader to get to the meat of the question and start matching the propositional makeup of the item with what is stored in memory. The prediction of item *p*-values by the semantic variables and jargon indicates that item difficulty results from the closeness of the match between the propositions in the item and propositions stored by the examinee from what was obtained through

803 BEST COPY AVAILABLE

the study reference material. Test items, then, should be syntactically simple and provide propositions as similar to those in the reference material as possible.

There is, however, more to test-item readability than just the item characteristics. It is possible that the examinee controls the majority of the variance in item readability and that the items themselves can be massaged only so much. More items with greater variability in their characteristics need to be analyzed to determine if this conclusion is valid.

Bibliography

Bloom, B.S., M.D. Engelhart, G.J. Furst, W.H. Hill, and D.R. Krathwohl, 1956. Taxonomy of educational objectives (subtitle: The classification of educational goals): *Handbook 1, The cognitive domain.* New York: David McKay Company, Inc. Portions reprinted in G.H. Bracht, K.D. Hopkins, and J.C. Stanley (eds.), 1972. *Perspectives in educational and psychological measurement.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., Selection 12.

Duncan, D. Emil. *A Theory and Model of Item Readability.* Published in the Minutes of the 23rd Annual Conference of the Military Testing Association, 1981.

Filmore, C. J. Some problems in case grammar. Cited in Kintsch, W., *The Representation of Meaning in Memory.* Hillsdale, N.J.: L. E. Earlbaum Associates, 1974.

Kincaid, J.P., Aagard, James A., and O'Hara, John W. *Development and Test of a Computer Readability Editing System (CRES).* Training Analysis and Evaluations Group Report No. 83, March, 1980. TAEG; Orlando, Florida, 321813.

Kintsch, W. *The representation of meaning in memory.* Hillsdale, N.J.: L. E. Earlbaum Associates, 1974.

Kintsch, W. and Keenan, J.M. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology,* 1973, 5, 257-274.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., and Keenan, J. M. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior,* 1975, 14, 196-214.

Kintsch, W. and Monk, D. Storage of complex information in memory: Some implications of the speed with which inferences can be made. *Journal of Experimental Psychology,* 1972, 94(1), 25-32.

Kintsch, W. and Vipond, D. Reading Comprehension and Readability in Educational Practice and Psychological Theory. In

Lambert, J.V. and Siegel, A.I. Psycholinguistic determinants of readability. In a.I. Siegel and J.R. Burkett (Eds.) *Application of structure-of-intellect and psycholinguistic concepts to reading comprehensibility measurement.* AFHRL-TR-74-49, Lowry AFB, Colo., 1974.

Madden, H.L. and Tupes, E.C. *Estimating reading ability level from the AQE General Aptitude Index.* Lackland Air Force Base, Texas: Personnel Research Laboratory, Aerospace Medical Division, PRL-TR-66-1, AD-632 182, February 1966.

Schwartz, D., Sparkman, J.P., and Deese, J. The process of understanding and judgments of comprehensibility. *Journal of Verbal Learning and Verbal Behavior,* 1970, 9, 87-93.

804

BEST COPY AVAILABLE

A COMPARISON OF SERVICE JOB STANDARDS FOR FOUR MILITARY SPECIALTIES

Brian K. Waters

Human Resources Research Organization

The U.S. Military Services (Army, Navy, Marine Corps, and Air Force) enlist over 300,000 new active duty recruits each year. The Services use numerous criteria, primarily aptitude composite scores, to assign a recruit to one of the hundreds of Service specialties.1 It would seem reasonable that aptitude entry requirements for similar jobs across Services would be somewhat similar. Armed Services Vocational Aptitude Battery (ASVAB) composite score entrance requirements for four different kinds of jobs across the four Military Services, (Radio Operator, Jet Engine Mechanic, Artilleryman, and Administrative Specialist) are analyzed in this paper. Entry standards for each specialty are shown as well as the proportion of the national youth population who meet minimum scores for each specialty/Service. Service recruiting goals for FY 1986 for these jobs, in terms of desired distributions of AFQT scores, are discussed along with their relationship with the minimum standards. Similarities and differences in aptitude level requirements across Services and specialties are examined for their implications for manpower policy.

SERVICE DIFFERENCES

It is important that the reader understand at the outset that the Military Services have distinctly different, Congressionally-mandated functions which affect the kinds of skills, knowledges, and abilities that they need to fulfill their respective missions. The Army and Marine Corps, for example, have extensive ground combat responsibilities which are quite different from most Air Force and Navy specialties which tend to be more technically oriented. Certainly the environment of a ship is very different from an aircraft, tank, or infantry foxhole. Even for what is ostensibly the "same" job, such as aircraft mechanic, the particular weapon system may dictate a quite different mix of abilities for job incumbents. These job differences, and the historical, cultural, and organizational traditions of the Services dictate that each one set its own classification standards. The Services have had long-standing research programs to validate their standards, generally based upon the statistical relationship between aptitude test scores and performance in military training. More recently, the Service personnel research laboratories have been trying to validate their selection and classification procedures against actual on-the-job performance -- a far more difficult, though also more desirable basis for the establishment of minimum aptitude standards.

---

1 A generic term "specialty" is used in this paper to denote occupations across Services, including Navy ratings.

2 Paper presented at the 27th annual meeting of the Military Testing Association in San Diego, CA, October 24, 1985. The opinions expressed in this paper are those of the author, and do not represent official DoD or Service policy.

SELECTION OF SPECIALTIES FOR ANALYSIS

It was desired that a sampling of military specialties be used to provide a broader examination of Service classification processes beyond a single job. Several criteria were used to select the specialties for analysis from the hundreds of Service specialties. First, candidate specialties had to be part of the 32 occupational specialties being investigated in the Joint-Service sponsored job performance measurement (JPM) project. Second, specialties with cross-Service equivalent jobs were sought. Third, a representation of mechanical, administrative, electrical, and general specialties was desired. Fourth, at least one combat specialty would be included, and finally, specialties which had a relatively large number of annual accessions in each Service would be most useful.

The DoD Occupational Conversion Table, which categorizes Service specialties into common DoD occupational groupings in terms of similar tasks, provided the basis for cross-Service specialty matches. Four DoD Occupational Codes were selected. Table 1 shows the four primary specialties and their equivalent jobs in the other Services.

TABLE 1 Occupational Specialties Selected

DoD Occupational Code (Type)	Specialty Data	Military Service			
		Army	Navy	Marine Corps	Air Force
041 (Combat)	Specialty	Cannon Crewman	Gunner's Mate	Field Artillery[b] Batteryman	[a]
	Service Code	13B	GM	0811	--
	#	24,925	7,462	2,821	--
201 (Electrical)	Specialty	Radio Teletype Operator	Radioman[b]	Field Radio Operator	Ground Radio Operator
	Service Code	05C	RM	253X	2935X
	#	8,691	16,181	5,109	1,749
510 (Administrative)	Specialty	Admin Specialist[b]	Yeoman	Admin Clerk	Admin Specialist
	Service Code	71L	YN	015X	702X0
	#	24,418	11,986	4,243	27,528
601 (Mechanical)	Specialty	Aircraft Power-Plant Repairman	Aviation Machinist's Mate	Aircraft Power Plant Mechanic	Jet Engine Mechanic[b]
	Service Code	68B	AD	602X	426X2
	#	840	13,340	1,122	10,620

[a]No Air Force equivalent specialty to DoD Occupational Code 041.
[b]Primary JPM project Service Specialty.

The four jobs selected include about nine percent of Army enlisted positions, ten percent of Navy billets, eight percent of Marine Corps jobs, and eight percent of Air Force enlisted positions. Although certainly not statistically representative of the hundreds of separate specialties in the four Services, they do suggest how a relatively large number of new recruits are matched with their jobs.

STUDY LIMITATIONS

The target population for this study is limited to the U. S. Military Services' active duty enlisted personnel policies in March, 1985. Information provided during interviews with Service and DoD policy personnel may not reflect official positions or policies. Data shown reflect only the four specialties examined and do not generalize to other occupations.

SPECIALTIES ANALYZED IN THIS STUDY

Table 1 depicts the four specialties chosen for analysis across the four Services. DoD has arrayed all Service specialties into categories which are similar to one-another based upon detailed task analyses of the jobs by occupational analysts. Clusters of the specialties analyzed in this study are identified by the DoD Occupational Codes shown on the far left of the table. Service-specific aptitude composites are made up of three or more ASVAB subtest scores as shown in Table 2. It should be noted that the Services use different score scales for their composites. Thus, comparisons of composite

TABLE 2 Minimum Entry Standards for Four Analyzed Specialties, by Service

DoD Code	Military Service	Specialty Title	Specialty Code	Classification Composite	Minimum Score	Percent PAY[a] Qualified
041	Army	Cannon Crewman	13B	AR+CS+MC+MK	85	71.3
	Navy	Gunner's Mate	GM	AR+EI+GS+MK	204	54.6
	Marines	Field Artillery Batteryman	0811	AR+CS+MC+MK	90	67.4
	Air Force	--	--	--	--	--
201	Army	Radio Teletype Operator	05C	AS+CS+NO+PC+WK	100	51.6
	Navy	Radioman	RM	AR+CS+NO+WK	149	55.7
	Marines	Field Radio Operator	253X	AR+EI+GS+MK	90	66.5
	Air Force	Ground Radio Operator	2935X	CS+NO+PC+WK	50	42.8
510	Army	Administrative Specialist	71L	CS+NO+PC+WK	95	56.0
	Navy	Yeoman	YN	AR+CS+NO+WK	165	30.4
	Marines	Administrative Clerk	015X	CS+NO+PC+WK	100	43.3
	Air Force	Administrative Specialist	702X0	CS+NO+PC+WK	35	54.1
601	Army	Aircraft Power Plant Repairman	68B	AS+EI+MC+NO	100	54.2
	Navy	Aviation Machinist's Mate	AD	AR+GS+EI+MK	190	64.4
	Marines	Aircraft Powerplant Mechanic	602X	AS+EI+MC+NO	100	54.1
	Air Force	Jet Engine Mechanic	426X2	AS+GS+MC	38	58.9

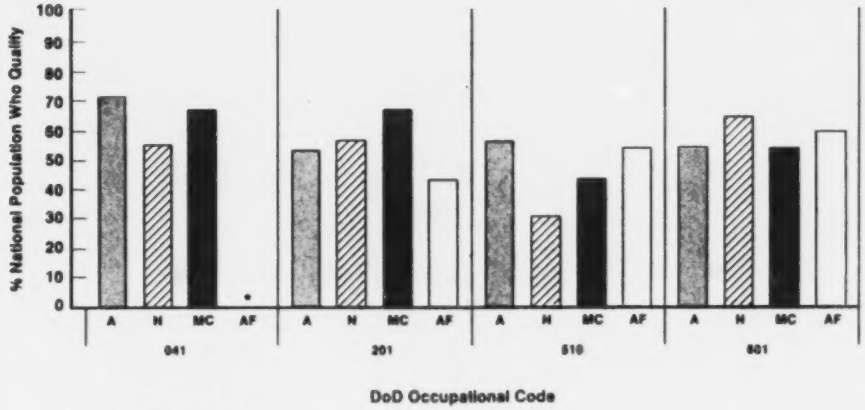
SOURCE: Eitelberg, M.J., with Lathrop, M.E., & Laurence, J.H. *Manpower for Military Occupations*, (1985). Naval Postgraduate School: Monterey, CA.

[a]18-23 year old males qualified for specialty (From Profile of American Youth Population)

NOTE: ASVAB Subtest abbreviations are as follows:

AR	=	Arithmetic Reasoning	MC	=	Mechanical Comprehension
AS	=	Automotive-Shop Information	MK	=	Mathematics Knowledge
CS	=	Coding Speed	NO	=	Numerical Operations
EI	=	Electronics Information	PC	=	Paragraph Comprehension
GS	=	General Science	WK	=	Word Knowledge

cutting scores should not be made across Services. Figure 1 displays the proportion of the 18-23 year old national population who meet minumum aptitude composite score standards for the 041, 201, 510, and 601 DoD Occupational Code specialties analyzed in this study.



*No Air Force Specialty in Code 041

SOURCE: Ellsberg, 1985

Figure 1. Percent of National Population of 18-23 Year Old Males Who Qualified for Service Entry and Specialty Entry by DoD Occupational Code and Service.

Table 2 and Figure 1 contain a great deal of data about the effects of formal Service standards on qualification for these four jobs. Differences within similar jobs across Services and across the four specialities are striking. For example, in the administrative specialties (DoD Code 510), 56 percent and 54 percent of youth would qualify for the Army 71L and Air Force 702X0 specialties respectively, but only 30 percent for the ostensibly comparable Navy rating of Yeoman. Either the jobs are not really similar as the DoD Occupational Code suggests, or other policy variables are entering the standard setting process. This type of inter-Service difference is common, and will likely complicate the cross-Service use of JPM project results. Similarly, looking at these four specialties across the four Services yields inconsistent results. The most difficult of the jobs to qualify for in the Navy is Yeoman; while for the Army, it is Radio Teletype Operator; for the Marines, Administrative Clerk; and for the Air Force, Ground Radio Operator. Of the four specialties examined, lowest aptitude composite scores are required to qualify for Cannon Crewman, Aviation Machinist's Mate, Field Artillery Batteryman and Field Radio Operator, and Jet Engine Mechanic for the four Services, respectively. Certainly there are quite different Service policies reflected in these data.

Table 3 displays the Services' AFQT category recruiting goals for the four specialties from a recent report to Congress. (DoD, 1985) A remarkable diversity exists between Services in their desired aptitude distributions for new recruits in ostensibly the same jobs. A comparison of these Service requirements is also surprising when considered along with the previously discussed qualification standards for these jobs. For example, the Navy goal is 100 percent of its Yeomen recruits to be in AFQT Categories I to IIIA (the upper half of the population). The implications of these data on recruiting costs and supply/demand are obvious, though the rationale appears less so. An excellent discussion of these issues is available in DoD, 1985, Volume I. Other similar comparisons of Tables 2 and 3 data are equally provocative. In some cases, it appears that goals may be driven by a Service's recruiting market rather than by Service personnel requirements.

TABLE 3 Military Service FY 1986 AFQT Recruiting Goals for Four Specialties

DoD Code	Military Service	Specialty Title	Specialty Code	FY 1986 Percent AFQT Category Goals I-IIIA	IIIB	IV
041	Army	Cannon Crewman	13B	58	30	12
	Navy	Gunner's Mate	GM	60	28	12
	Marines	Field Artillery Batteryman	0811	54	37	9
	Air Force	--	--	--	--	--
201	Army	Radio Teletype Operator	05C	64	27	9
	Navy	Radioman	RM	32	48	20
	Marines	Field Radio Operator	253X	61	36	3
	Air Force	Ground Radio Operator	2935X	90	10	0
510	Army	Administrative Specialist	71L	64	29	7
	Navy	Yeoman	YN	100	0	0
	Marines	Administrative Clerk	015X	71	28	1
	Air Force	Administrative Specialist	702X0	90	10	0
601	Army	Aircraft Powerplant Repairman	68B	70	24	6
	Navy	Aviation Machinist's Mate	AD	73	20	8
	Marines	Aircraft Powerplant Mechanic	602X	74	26	0
	Air Force	Jet Engine Mechanic	426X2	79	21	0

SOURCE: DoD, 1985.

CONCLUSION

Although it is clear that the individual Services must retain the prerogative of setting their own classification standards based upon their own personnel requirements, the analysis in this paper opens many questions in the mind of the military manpower analyst. Why are there such diverse entry standards for what are ostensibly the same jobs across Services? Why do the Services require radically different aptitude distributions for personnel fulfilling the specialties with the same DoD Occupational Codes?

The questions raised seem to point toward the methodologies used by the Services to define their personnel requirements. As shown by their submissions to the Congress (DoD, 1985, Vol. 1), their methods for determining their manpower requirements are not consistent. There appears to be a high priority need for research into establishing a methodology for determining the aptitude needs of the Services based upon a defensible way of specifying minimum requirements for each specialty. Such a methodological breakthrough would increase the efficiency of the assignment of personnel to specialties in all of the Services, and would provide a sound basis for requests for recruiting and training resources to the Congress.

REFERENCES

Department of Defense
1985 Report to the House and Senate Committees on Armed Services: Defense Manpower Quality, Volumes I, II, & III. Washington, D.C.: Office of the Assistant Secretary of Defense (Manpower, Installations, and Logistics).

Eitelberg, M. J., with Lathrop, M. E., and Laurence, J. H. Manpower for
1985 Military Occupations. Monterey, CA: Naval Post Graduate School.

EXPLORING A STATISTICALLY VIABLE ASSIGNMENT BASIS USING ASVAB

DR. M. MARK SCHWARTZ

U.S. ARMY INTELLIGENCE SCHOOL
FT. DEVENS, MA

INTRODUCTION. The utilization of the ASVAB tests for determining likelihood of success in MOS training is premised on establishing minimum scores on 1 or 2 of the tests, above which a person is deemed suitable for training in that MOS. This minimum-score concept, in essence, states that (a) all persons above this score are equally likely to succeed and (b) the other tests in the battery are not essential information for determining success in MOS training.

These two points raise issues of (a) is there a maximum score above which a person may be not qualified, and (b) why not use all the data collected? Although there is a range of personal and training-environment factors influencing the likelihood of training success, this paper addresses only the characteristics captured and described by the ASVAB tests. The test results are aptitude/ability indices, summarized in a profile of 10 scores.

Sample profiles are:

TABLE I

	CO	FA	EL	OF	GM	MM	CL	ST	GT	SC
STUDENT 1	135	147	135	124	128	122	141	138	133	133
STUDENT 2	89	94	108	88	109	95	92	101	95	91

As an introduction to this paper's direction, these are real data, where student 1 did not succeed (DROP) and student 2 did succeed (GRAD). Using these two for discussion purposes, these could be called two good examples of "other factor" influence. Both met the criterion for entrance, although the seemingly better candidate wasn't successful. An alternate reaction to these two cases could be that this is simply an example of predictable and expected" classification error". However, both of these explanations dismiss the possibility that there is a meaningful source of explanation in the context of the available data. Even a response of "oh my, it's time to revalidate", although a professional response, is not the only alternative. All three of these reactions, I contend, divert our attention from the system, which collects a lot of data, summarizes it in a profile of 10 scores, and then proceeds to ignore 80 to 90% of the available information in the decision-making process. This is not a condemnation but rather a description, and it begs the question: does the data contain useful patterns of information?

FOCUSING THE QUESTION. By definition, the word "patterns" states that the entire profile and aggregates of profiles are to be addressed. This is immediately different from the concept of selecting one or two scales from the set and then fixing a point on those scales as criteria. The assessment schema has been shifted from looking at individual scores on an absolute scale to looking at the entire profile collectively, and on a relative scale.

Just as the word "patterns" did, the word "relative" signals a different conception. In doing this, we are allowing that there may be contributory and interacting effects of the characteristics these 10 test scores describe. In essence, let us assume that all the information is relevant and that what needs to be determined is how much of each one, relatively, is present in GRADS and DROPS. For example, referring to the two sample profiles in Table 1, the focus is now on the relative amounts of ability/aptitude each person possesses. Notice that on an absolute scale, case 1 and case 2 are distinctly different if we compare their scores on any given test, and although student 1 has higher scores on every test, both students show different patterns of ability/aptitude within their own profile. One way to easily see this is to rank each individual's score profile--in essence, for each individual what are the relative amounts of CO, FA, EL, ETC,-- and in doing so, a picture emerges which shows discrimination between these students, but in a different context.

Ranked, with ties, the two profiles become:

TABLE II

TEST	1	2	3	4	5	6	7	8	9	10
DROP	6.5	10	6.5	2	3	1	9	8	4.5	4.5
GRAD	2	5	9	1	10	6.5	4	8	6.5	3

Notice that tests 4,8,9, and 10 show roughly equivalent ranks (both high or low), while tests 3,5, and 6 are low for the DROP when compared to the GRAD, and conversely for tests 1,2, and 7. Note that test 8, which is used for the school selection criterion, shows an "8" for both individuals, indicating that within each individuals repetoire of skills/knowledge/ability/aptitude/-etc., the characteristics captured in this test rank the same relative to the other characteristics in the profile. This analysis presents a different picture from the one seen when looking at raw ASVAB scores. Table I shows clear discrimination between DROP and GRAD, while Table II indicates three different nodes of information -- one in favor of the DROP, one in favor of the GRAD, and one "even". What is now occuring is a horizontal and vertical scanning of"quantity of test X", and it is this simultaneous within and between comparison that provides a basis for "collective and relative" comparison. The question now is

whether this is a significant and useable pattern within and between groups of DROPs and GRADs, and if so, how can these patterns be assessed?

PSYCHOMETRIC CROSSFERTILIZATION. Ranking ignores absolute ASVAB value, as in the case of the ranks of test 4, where a score of 124 (rank = 2) is roughly equivalent to a score of 88 (rank = 1). Realizing inherent theoretical questions about rank and the use of rank means, in order to step up from talking about individuals to groups the mean is the most convenient metric to use. Then each individual rank can be compared to the group GRAD and DROP mean for each test to see if on that characteristic, the individual "looks more like" A DROP or a GRAD. The expression "looks more like" translates roughly to does the individual look more like a member of their group or the other group, using closer-to-the-mean as the criterion.

The purpose of this comparison is to transform the individual's profile of ranks, which was tranformed from the ASVAB profile, into a profile of yes-or-no, symbolized as 1 = looks more like a GRAD on this test, and 0 = looks more like a DROP on this test.

Below is a summary of all the transformations so far.

TABLE III

	ASVAB		Rank		Group Rank Means		Status (G=1; D=0)		Overall Classification	
	Drop	Grad	Drop	Grad	Drop	Grad	Drop	Grad	Drop	Grad
TEST										
1	135	89	6.5	2	5.17	5.05	0	1	Drop	Grad
2	147	74	10	5	7.17	6.35	0	1		
3	135	108	6.5	9	5.53	5.17	0	1		
4	124	88	2	1	4.83	4.63	1	1		
5	128	109	3	10	4.17	4.98	0	1		
6	122	95	1	6.5	3.23	3.60	0	1		
7	141	92	9	4	6.48	6.05	0	1		
8	138	101	8	8	5.85	6.57	1	1		
9	133	95	4.5	6.5	7.03	6.70	1	1		
10	133	91	4.5	3	5.53	5.90	0	0		

The status profiles can be examined to assess an individual's standing on each test. The question is how to collectively summarize all these individual assessments. It is to be noted that performing these transformations can contribute to some loss of fidelity. However, the ultimate results indicate that a viable system is operating. The answer to the question of how to collectively assess the overall status from the individual status indices lies in the application of a Bayesian technique for calculating the probability of group membership. Item Response Theory relies on the natural dichotomy of an item being right or wrong, plus other important item characteristics. But the

concern here is not right or wrong, but "looking like". This Bayesian procedure takes into account each groups total "likeness" for each test (something like the p-value in item analysis) and uses this information with the individual's "likeness" to compute a probability, using all 10 scores, that the individual is a GRAD or DROP. Despite the apparent outcome of classifying the GRAD and the DROP when looking at the status column in Table III, it is to be noted that these cases were selected for demonstration, and not all cases are as clear.

AN EMPIRICAL VIEW AND DISCUSSION. Two simple tests of this conception were done. The procedure was to take 30 actual DROPS and GRADS ASVAB data (a second case had 16 each), transform the scores, apply the Bayesian procedure, and then compare the actual status with the classification outcomes.

The results were:

TABLE IV

		Classification				Classification	
		GRAD	DROP			GRAD	DROP
ACTUAL	GRAD (N=30)	21	9	GRAD (N=16)	12	4
	DROP (N=30)	11	19	GRAD (N=16)	6	10
	TOT=	32	28	TOT=	18	14
		Case 1			Case 2	

Statistically, both cases are significant, with chi-square = 6.70 for case 1 and 8.57 for case 2. Classification error for both cases are:

TABLE V

	Case 1	Case 2
Grad Error Rate	39%	25%
Drop Error Rate	37%	38%

The values in Table V are roughly equivalent, which can be interpreted as an indication that the approach used operates across these 2 MOSs, suggesting that there may be a viable basis for expectation that it can be applied for all MOS.

Given the small samples used in this pilot research, the somewhat high error rates are just indicators and neither a minimum nor a maximum expectation. As sample size increases, there is no certainty as to the expected change in mean rank, which is a key

value in the overall process. Theoretically, if the law of large numbers operates, all means would go to 5.5. However, some of the reported means in Table III suggest that the likelihood is that a systematic, non-random principle is operating, and the means would not converge.

CONCLUSION. This is by no means conclusive, but rather exploratory. The goal here was not to have error-free classification but rather to demonstrate that models other than the current ones used for assessing profiles of data do show promise and merit further research. This model could be extended to the point of realizing that all the requisite data could be compiled and updated at each school, transmitted to every recruiting office, and utilized to evaluate the likelihood of successful training in a range of MOSs by simply entering an individual's profile of scores into a computer.

The Validity of ASVAB for Predicting Training and SQT Performance

Paul G. Rossmeissl
U.S. Army Research Institute[1]

Donald H. McLaughlin, Laurese L. Wise and David A. Brandt
American Institutes for Research

This paper is a condensation of a larger report (McLaughlin, Rossmeissl, Wise, Brandt, & Wang; 1984) which investigated the validity of the Armed Services Vocational Aptitude Battery (ASVAB) for predicting success in Army jobs or Military Occupational Specialties (MOS). The ASVAB is a cognitive test battery used by the military services as their primary instrument for selecting and classifying enlisted personnel. This particular research was based upon ASVAB forms 8/9/10 which was composed of ten subtests: General Sciences (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto/Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehensive (MC), and Electronics Information (EI). The two verbal subtests, WK and PC, are most often combined into a single measure of verbal ability called VE. The current version of ASVAB (forms 11/12/13) uses parallel forms of these same subtests.

Scores on the ten ASVAB subtests are typically combined into aptitude area (AA) composites. Examples of these composites are given in Table 1. The Army composites serve as the basis for assignment of personnel to Army MOS in that a minimum qualifying score on one of the aptitude area composites is required for admission to Army initial level training courses. For example, the CO composite is used to classify recruits into the infantry and armor specialties. Similarly, the MAGE composites are used by the Air Force to select and classify prospective personnel into Air Force specialties. The final set of composites routinely in use are the High School Composites which have been developed for use when ASVAB is administered to high school students as a career guidance tool. Maier and Truss (1983) have also recommended that the first four of these composites be used to select and classify enlisted personnel within the Marine Corps.

The goal of the McLaughlin et al. (1984) research was twofold. First, the validities of the composites then in use by the Army and other DoD agencies were evaluated with regard to predicting success within the Army. Second, an additional set of composites were derived empirically in hopes of obtaining a composite system with maximal predictive validity.

In all cases the validation criterion were MOS specific end-of-course training scores or skill qualification tests (SQTs). All of the criterion measures were trimmed of outliers and then standardized before any

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

validation analyses. The separate training and SQT data were combined for validation analyses at the MOS level. All validities were corrected for restriction of range using the multivariate adjustment due to Lawley (1943) and described by Lord and Novick (1968).

Table 1
Typical ASVAB Composites

Army Composites (1983)		
Clerical/Administrative	CL	VE + NO + CS
Combat	CO	AR + CS + AS + MC
Electronics Repair	EL	GS + AR + MK + EI
Field Artillery	FA	AR + CS + MK + MC
General Maintenance	GM	GS + AS + MK + EI
Mechanical Maintenance	MM	NO + AS + MC + EI
Operators/Food	OF	VE + NO + AS + MC
Surveillance/Communications	SC	VE + NO + AS + CS
Skilled Technical	ST	VE + GS + MK + MC
MAGE Composites		
Mechanical	M	MC + AS + GS
Administrative	A	VE + NO + CS
General	G	AR + VE
Electronic	E	AR + MK + GS + EI
High School Composites		
Mechanical Trades	HSMT	AR + MC + AS + EI
Office and Supply	HSOS	VE + CS + MK
Electronics/Electrical	HSEE	AR + EI + MK + GS
Skilled Services	HSSS	AR + VE + MC
Academic Ability	HSAA	AR + VE

Composite System Validities

Table 2 gives the adjusted validities for each of the composite systems displayed in Table 1. Validities and sample sizes are given for each of the nine clusters of MOS now in use by the Army. The validities were obtained by averaging the validities for the individual MOS within each cluster and weighting by the number of soldiers within each MOS.

Table 2
Validities of Established Composite Systems

Army Composites (1983)

Cluster of MOS	(N)	Composite CL	CO	EL	FA	GM	MM	OF	SC	ST
CL	10368	48	51	53	54	49	46	50	50	53
CO	14266	36	44	43	43	43	42	44	40	44
EL	5533	38	47	46	47	46	47	44	44	47
FA	5602	39	49	48	48	49	49	49	45	44
GM	2571	39	48	46	46	47	48	48	45	47
MM	7073	36	48	46	45	48	48	48	43	46
OF	8704	38	48	47	45	48	47	48	44	48
SC	3729	39	49	48	47	48	47	48	45	49
ST	7061	51	56	57	57	55	54	56	54	58

MAGE Composites

Cluster of MOS	(N)	Composite M	A	G	E
CL	10368	45	48	54	53
CO	14266	42	36	42	43
EL	5533	45	38	46	47
FA	5602	48	39	46	48
GM	2571	46	39	44	46
MM	7073	48	36	44	46
OF	8704	47	38	47	47
SC	3729	47	39	47	48
ST	7061	52	51	57	57

High School Composites

Cluster of MOS	(N)	Composite HSAA	HSMT	HSOS	HSSS	HSEE
CL	10368	54	47	54	53	53
CO	14266	42	43	40	44	43
EL	5533	46	47	43	47	47
FA	5602	46	49	44	49	48
GM	2571	44	47	43	47	46
MM	7073	44	49	41	47	46
OF	8704	47	48	43	48	47
SC	3729	47	48	44	49	48
ST	7061	57	54	56	58	57

732 S1S

The main diagonal of the upper portion of Table 2 gives the validities of the composites that were associated with each of the nine clusters in 1983. The most interesting feature of the data in Table 2 is the uniformity of the validities. All of the entries are between .36 and .58, with the mean validity of each system being about .45. One MOS cluster, ST, appears to be slightly more predictable than the others; and another cluster, CO, appears to be slightly less predictable. The remaining clusters show very little variance.

Identification and Validation of Alternative Composites

In order to develop alternative composites the MOS were partitioned into clusters, based on similarity of ASVAB profiles of successful criterion performance. The similarity between each pair of cells was defined as correlation of the predicted criterion performance in the two cells for the applicant sample. The performance predictions were based on ridge regressions, using the ASVAB subtests as predictors. The cells were clustered by adapting standard "leaf to stem" procedures. Upon finding that the results of the clustering were unstable, due to the high intercorrelations of the predicted criterion scores, the clustering procedure was modified to use as a starting point the Army's current grouping of MOS into aptitude area clusters.

Once a cluster had been defined the unit-weight composite with maximal predictive validity for that cluster was identified. It was found that optimal unit-weight composites for four clusters possessed a root mean square (RMS) predictive validity within 97% of the RMS validity of the ridge regression vectors computed separately for each of the 98 MOS included in the sample. The composition of these four alternative composites are given in Table 3, and their predictive validities are given in Table 4.

Table 3
Optimal Four Composite Solution

Composite		Subtests
Clerical/Administrative	(ACL)	VE + AR + MK
Skilled Technical	(AST)	VE + AR + MK + AS
Operations	(AOP)	VE + AR + MC + AS
Combat	(ACO)	VE + MK + MC + AS

Inspection of Table 4 shows that by focusing on the most valid portion of the ASVAB, the primary aim of this aspect of the research was achieved: the validities went up. The aggregate RMS predictive validity for the four alternative composites for their assigned MOS is .486, in comparison with RMS validity for the 1983 Army composites of .454. Certain members of the 1983 Army composite set account for a large part of the difference in validity between the two composite sets. When compared

Table 4
Predictive Validities of the Alternative Composites

Cluster of MOS	(N)	Composite ACL	AST	ACO	AOP
CL/ACL	10368	56	54	52	51
CO/ACO	14266	42	44	44	44
EL/ACO	5533	46	48	48	48
FA/ACO	5602	47	49	50	50
GM/ACO	2571	45	48	48	48
MM/AOP	7073	44	48	49	49
OF/AOP	8704	46	49	49	49
SC/AOP	3729	47	49	50	50
ST/AST	7061	58	58	57	57

to validities of the optimal composites for the same cluster of MOS, the 1983 Clerical composite (CL) appeared to be weak, with a validity of .48 versus a potential of .56. Another composite Surveillance and Communications (SC), was mildly weak, with a validity of .45 versus a potential .50.

Recommendations

A major purpose behind the McLaughlin et al. (1984) report was to present recommendations to the Army as to how the composite system then in use to select and classify enlisted personnel could be improved. The average validity of the set of four empirically derived alternative composites was .48 versus .45 for the existing composite systems. Thus, from a purely statistical point of view the results in terms of predictive validity tended to favor the alternative four composite solution over the nine composite system the being used or any of the alternatives being used by other armed services.

However, considering the costs of implementing a whole new composite system, it was decided that a more favorable proposal would be to maintain a nine composite system but to replace the the two composites which were the major source of the deficiency of the 1983 composites. The new CL composite would be comprised of the VE, AR, and MK subtests and would have a predictive validity of .56. The new SC composite would have a predictive vaidity of .50 and be made up of the VE, AR, MC, and AS subtests. The average validity of the revised nine composite system would be .47. The Army officially adopted this composite system on October 1, 1984.

The gain in expected performance resulting from the change in the CL and SC composites can only be approximated, because of the constrained nature of the selection and classification process. If, however, the choice were purely between assignment to an individual MOS and rejection, application of Cronbach's formula yields an expected gain of .05 standard deviations of criterion performance per person in the two clusters of MOS from the introduction of the two revised composites.

References

Lawley, D. (1943). A note on Karl Pearson's selection formulae. Royal Society of Edinurgh, Proceedings, Section A. 62, 28-30.

Lord, P., & Novick, M. (1968). Statistical theory of mental test scores. Reading MA: Addison-Wesley Publishing Company, Inc.

Maier, M. H., & Truss, A. R. (1983). Validity of ASVAB Forms 8, 9, and 10 for Marine Corps training courses: Subtests and current composites. (Center for Naval Analyses Memorandum No. 83-3107). Alexandria, VA: Marine Corps Operations Analysis Group.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A. & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY 1981 and FY 1982 enlisted accessions. Technical Report No. 651, U. S. Army Research Institute for the Behavioral and Social Sciences, Alexandria VA.

GENDER, ETHNIC GROUP, APTITUDE AND PERSONALITY DETERMINANTS OF U. S. COAST GUARD ATTRITION

ROBERT L. FREY, JR.

U. S. COAST GUARD HEADQUARTERS

All five Armed Services have been using aptitude measures for enlistment screening and technical school qualification for decades. Research consistently has shown that premature first term attrition (i.e., before the end of one's enlistment) is related to aptitude.

However, as we all know, aptitude is only a small part of the potential predictor space. In the Coast Guard we are interested in the possible use of personality assessment to improve our screening. One of the major responsibilities of the Coast Guard is to board vessels for fishing law inspections and drug interdiction investigations. These boardings are conducted by armed personnel; obviously, management is quite concerned that assessment of personnel to determine suitability for armed boarding party duty be as accurate as possible.

The nature of these boardings is similar in many respects to law enforcement activities. For example, it is critical that the people involved exercise utmost discipline and be able to deal with sudden stress without resorting to the use of firearms unless absolutely necessary. Naturally, extensive special training is given to personnel before thay are allowed to take part in armed boarding parties in order to meet such performance requirements. In addition, however, this is a performance domain where we expect people's behavioral predispositions to be quite predictive. Accordingly, we are investigating the potential of using personality profiles as part of the assessment process. A major objective in our use of personality profiles is to improve assessments of new recruits for their general suitability to serve and remain in the U. S. Coast Guard.

The personality assessment instruments which we used for this study were the Sixteen Personality Factor Questionnaire (16PF) and the Clinical Analysis Questionnaire (CAQ). There were three major reasons for our choice: (1) The 16PF/CAQ covers the full realm of both normal traits and clinical syndromes, 2) the 16PF/CAQ seems to be the best validated instrument for law enforcement and other occupations involving great stress, and 3) normative data show that the 16PF/CAQ profiles are virtually the same for minorities as for whites--that is, the 16PF/CAQ is fair for minorities.

RESULTS AND DISCUSSION

The subjects for this study were the cohort of FY83 recruits. The total N was 4,288. There were 3,544 whites and 744 minorities. Also, there were 3,732 males and 556 females. On the predictor side, aptitude test scores and 16PF/CAQ scores were obtained for each recruit. The aptitude measure used was a composite of verbal and arithmetic subtests from the ASVAB or a similar test battery (During part of FY83, the USCG was still using an old Navy battery). The aptitude composite has a normative mean of 100 and a standard deviation of 20. 16PF/CAQ scores are scaled to have a normative mean of 5.5, SD of 2 and a range of 1 to 10.

The analysis design was a crossed factorial multivariate analysis of variance. The three factors were: 1) Gender (male vs female), 2) Ethnic Group (Minority vs Majority), and 3) Status (In the CG vs out of the CG). Since the Coast Guard only has four year enlistments, anyone already out is a case of premature first term attrition. This resulted in a 2 X 2 X 2 design with eight cells. The criteria for this analysis were the aptitude score and the 16PF/CAQ scores.

Since assignment to the cells of the design is by self-selection, the cell N's were quite disparate. Accordingly, all the tests for interactions and main effects were accomplished using a hierarchical, least-squares model. The attrition rates by ethnic group were: 1) majority group - 16.4%, 2) minority group - 28.9%. The attrition rates by gender were: 1) males - 17.5%, 2) females - 25.7%. The overall attrition rate was 18.6%.

The Gender (G) X Ethnic Group (E) X Status (S) multivariate interaction was not significant ($F = 0.928$, $p < .575$). However, the E X S multivariate interaction was significant ($F = 1.692$, $p < .013$, $R = .105$). In order to interpret this multivariate interaction, the discriminant function structure coefficients (or loadings) were evaluated. The structure coefficients are used to interpret the underlying dimensionality of the E X S discriminant function. Four of the criteria explained the discriminant function:

CRITERION	LOADING
WARMTH	.501
APTITUDE	.499
LOW ENERGY DEPRESSION	- .318
SUICIDAL DEPRESSION	- .311

As defined in Krug, 1980

Warmth: High-scoring individuals are "...personable and easy to get along with. ...prefer to adapt to other people's schedules. ...share their feelings with others.

Low Energy Depression: "High scoring individuals report frequent feelings of sadness and gloom. ...almost never sleep soundly or wake up full of energy...have little zest for life and are worn out and low."

Suicidal Depression: "Centers around thoughts of self-destruction. High scoring individuals report that they are disgusted with life...entertain thoughts of death as a viable alternative to their present hopeless situation."

The univariate interaction means for these four criteria are:

	GROUP			
CRITERION	MAJ-IN	MAJ-OUT	MIN-IN	MIN-OUT
WARMTH	4.36	4.34	5.02	5.54
APTITUDE	108.2	104.3	99.2	98.8
LOW ENERGY DEPRESSION	5.01	5.77	5.05	5.48
SUICIDAL DEPRESSION	4.61	5.27	4.84	5.13

NOTE:

MAJ-IN: Majority group personnel still in the Coast Guard

MAJ-OUT: Majority group personnel already out of the Coast Guard
MIN-IN: Minority group personnel still in the Coast Guard
MIN-OUT: Minority group personnel already out of the Coast Guard

By definition of an interaction, we would expect the criteria to operate differently within each ethnic group. For example, within the majority group, there is no difference in Warmth between those still in the Coast Guard vs those already out of the Coast Guard. In contrast, within the minority group, those who are already out of the Coast Guard scored considerably higher on Warmth than those who are still in the Coast Guard. Looking at the Aptitude score, within the majority group, those who are still in the Coast Guard scored considerably higher than those already out of the Coast Guard. Within the minority group, there is no difference on the Aptitude score between those still in the Coast Guard vs those already out of the Coast Guard. For both the Low Energy Depression score and the Suicidal Depression score, the interaction pattern is the same. That is, within the majority group, those who are already out of the Coast Guard scored considerably higher (i.e., "worse") than those who are still in the Coast Guard. Within the minority group, however, the difference between those still in the Coast Guard vs those already out of the Coast Guard is very small.

So far, we have examined those criteria which "explain" the E X S multivariate interaction. However, follow-on analyses are required to clarify the differential status profiles within each ethnic group. Separate analyses of the status factor (using the common multivariate error matrix) were done within each Ethnic Group. That is, an internal analysis of the E X S multivariate interaction was accomplished. The primary interest in the internal analysis was to see whether the discriminant function composite of aptitude and personality variables for predicting status was different for the majority and minority ethnic groups.

The results of the internal analysis indeed showed that the underlying dimensionality of the status prediction composite was quite different for the two ethnic groups.

The multivariate test of the status factor was significant within the majority group ($F = 8.4$, $P < .001$, $R = .229$). Five of the criteria explained the discriminant function:

CRITERION	LOADING
HYPOCHONDRIASIS	- .646
LOW ENERGY DEPRESSION	- .601
APTITUDE	.582
SUICIDAL DEPRESSION	- .559
PSYCHOLOGICAL INADEQUACY	- .534

As defined in Krug, 1980:

Hypochondriasis: "High-scoring individuals are depressed and preoccupied with bodily disfunctions. ...[they] feel that their health is worse than others ... feel sluggish ... and generally run down".

Low Energy Depression: (previously defined)

Suicidal Depression: (previously defined)

Psychological Inadequacy: "High-scoring individuals describe themselves as no good for anything ... think of themselves as doomed or condemned ... learned helplessness pattern".

The majority group univariate means for these criteria were:

	GROUP	
CRITERION	IN CG	OUT CG
HYPOCHONDRIASIS	4.69	5.47
LOW ENERGY DEPRESSION	5.02	5.72
APTITUDE	108.33	103.89
SUICIDAL DEPRESSION	4.60	5.31
PSYCHOLOGICAL INADEQUACY	4.88	5.51

The four personality scores noted above are clinical syndromes from the CAQ portion of the 16PF/CAQ. Thus, a high score is "worse" than a low score. As expected, those majority group members already out of the Coast Guard scored higher than those still in the Coast Guard. Depression and feelings of inadequacy seem to be a common theme for the majority group members already out of the Coast Guard. Also, those already out of the Coast Guard had an aptitude score lower than those still in the Coast Guard.

The multivariate test of the status factor was significant within the minority group ($F = 2.034$, $P < .001$, $R = .115$). Five of the criteria explained the discriminant function:

CRITERION	LOADING
WARMTH	- .492
HYPOCHONDRIASIS	- .490
SENSITIVITY	- .462
ANXIOUS DEPRESSION	.395
PSYCHOLOGICAL INADEQUACY	- .390

As defined in Krug, 1980:

Warmth: (previously defined)

Hypochondriasis: (previously defined)

Sensitivity: High scoring individuals are "... tender-minded, dependent, overprotected, fidgity, clinging, and insecure. ... They prefer to use reason rather than force in getting things done".

Anxious Depression: "High-scoring individuals describe themselves as clumsy and shakey ... they lack self-confidence, and seldom speak out and say what they think. They are confused and unable to cope with sudden demands..."

Psychological Inadequacy: (previously defined)

The minority group univariate means for these criteria were:

	GROUP	
CRITERION	IN CG	OUT CG
WARMTH	5.02	5.53
HYPOCHONDRIASIS	4.85	5.37
SENSITIVITY	5.19	5.71
ANXIOUS DEPRESSION	5.32	5.76
PSYCHOLOGICAL INADEQUACY	4.88	5.51

It is immediatley apparent that the dimensionality of the discriminant function composite for the minority group is different from that of the majority group. First of all, aptitude is not involved. Also, two normal traits from the 16PF portion of the test are part of the composite (Warmth and Sensitivity). Interestingly, those minority group members already out of the Coast Guard score higher on these 2 criteria than do those still in the Coast Guard. As expected, those minority group members already out of the Coast Guard score higher on Hypochondriasís, Anxious Depression and Psychological Inadequacy than do those still in the Coast Guard.

Regarding the Warmth and Sensitivity criteria, minorities scored higher than the majority. Since the USCG enlisted force is composed of 83% majority members, one could posit that the majority group culture dominates the enlisted force. It may be that there is a "cultural clash" in the USCG for a significant number of the minority recruits. Thus, the higher Warmth and Sensitivity scores of the minority group members already out of the Coast Guard may indicate those who were most discrepant from the dominant culture.

There is also a practical significance to the separate discriminant function composites for majority and minority group members. The status discriminant function composite was computed using two different methods: 1) based on the total sample, ignoring the E X S multivariate interaction (i.e., from the status main effect test), and 2) separately for each group, based on the internal analysis of the E X S multivariate interaction. Point biserial correlations between the discriminant function composites and actual status were then computed for each ethnic group. For the majority group members, there is no real change--.24 under method 1 and .25 under method 2. This is not surprising since the total sample is 82.6% majority members. However, for the minority group members the correlation under method 1 is .17 while the correlation under method 2 is .26. This is a meaningful improvement for screening purposes.

The results of this study indicate that when one expands the predictor domain beyond aptitude, differential prediction for majority ad minority ethnic groups may be called for. Naturally, in the wider context, the results of one study are always tentative. Meta-analyses (such as are now done with aptitude scores) will be necessary before any conclusions can be reached. With all the work being done on new and supplementary predictors, the possibility of different prediction dimensions for majority and minority group members should not be overlooked.

REFERENCE

Krug, S E Clincal Analysis Questionnaire Manual. Champaign, ILL.:IPAT, 1980

STUDY OF WASTAGE FROM THE TERRITORIAL ARMY (UK ARMY RESERVE)

D M Blyth

Personnel Psychology Division
Army Personnel Research Establishment
Ministry of Defence
Farnborough
Hampshire
United Kingdom

INTRODUCTION

1. At last year's MTA Conference the results of the first phase of a study of the wastage problem in the UK Army Reserve Forces were reported in a paper entitled "Retention in the UK Territorial Army". (Dennison and Blyth, 1984). At that stage the work consisted of a literature review which focussed on international comparisons of reserve force attrition; an analysis of the available wastage and turnover statistics; a programme of interviews and group discussions with serving reservists at a number of Territorial Army (TA) units. The interviews and group discussions were carried out by staff from the Army Personnel Research Establishment (APRE).

2. The purpose of this paper is to summarise the findings of the study as a whole including the second phase which consisted of a nationwide interview survey of people who had been discharged from the TA in the preceeding twelve months. This work was carried out by a commercial Market Research agency using a structured interview schedule designed in collaboration with APRE. The aims of the survey were to quantify the relative importance of the factors associated with wastage, to isolate those factors on which the TA could realistically be expected to take action, either locally or centrally, and to make practicable recommendations for improvements.

BACKGROUND

3. The British Army's reserves fall into two categories. The first consists of officers and soldiers who have completed their service with the Regular Army. These "Regular Reservists" have a compulsory liability for reserve service on mobilisation and moves are currently underway to encourage them to undertake a limited amount of peacetime training. They are not, however obliged to do so.

4. The second category consists of civilian volunteers (including some ex-regular soldiers and officers) who undertake military training in their spare time and comprise the Territorial Army.

5. The role of the TA is to provide units, some highly specialised, to reinforce the Regular Army at home and abroad in the event of mobilisation. The TA provides almost 30% of the Army's mobilised strength.

6. In 1981 the Ministry of Defence published plans to expand the TA in two phases from 70,000 to 86,000 by 1990 to meet NATO and other commitments. Actual increases in strengths, however, have not kept pace with increases in target strengths. For example, the average total strength of officers and soldiers in September 1984 was 72,041 (having fallen from a high point of 72,703 in April 1983) compared to the target figure of 75,051.

7. The urgent need to improve retention of TA personnel led to the commissioning by the then Director, Territorial Army and Cadets of a study of the reasons for high wastage to be carried out by APRE. The study consisted of an element that was carried out by APRE staff and a separate element for which a research agency with a field force of interviewers was used. The latter consisted of interviews with 426 ex-TA soldiers drawn from a cross section of regions and types of TA unit. The APRE element led to a number of conclusions and recommendations some of which were reported in the MTA Paper referred to earlier. However, rather than report separately the findings of the national survey of leavers, all of the main findings and conclusions from the study as a whole will be reported briefly.

RESULTS AND CONCLUSIONS

8. The question of whether the current 30% annual turnover is an unreasonably high figure or is capable of being improved upon significantly was examined. The element of turnover attributable to voluntary wastage cannot be measured accurately at present. This is due to inconsistencies among units in categorising leavers and in some cases, ignorance of the real reasons for leaving. It would be more difficult still to estimate how much voluntary wastage could be prevented by remedial action by the TA. However, in view of the similar turnover rates experienced by other countries' reserve forces and the consistency of our own turnover over the years, despite major improvements to terms and conditions, we conclude that no more than a modest improvements can be expected from action likely to be at the disposal of the TA.

9. The available statistics on strengths, wastage and turnover can be misleading. Strength figures include a high proportion of soldiers who have done no training nor attended camp for at least a year but have not been discharged. Discharge categories used to describe types of leaver are not used consistently across units nor within units. It was recommended that more informative monthly, quarterly and annual statistical breakdowns should be produced and that firm guidance should be given to units on the use of discharge categories.

10. The results of the Market Research survey were in broad agreement with those of the APRE interviews and group discussions. In the survey of leavers, when asked for the single most important reason for leaving the TA, 33% gave job pressures, including a change to shift work and related financial problems; 23% gave the harmful effect that the demands on time had on family and social life; 27% referred to inadequacies within the TA such as lack of action and boredom; 7% gave physical demands or medical reasons; the rest gave an assortment of other reasons.

Table 1 gives a breakdown of responses to the questions "What is the single most important reason for not carrying on with the TA?" and "What other things helped you to make up your mind?"

TABLE 1. REASONS GIVEN FOR LEAVING

(What is) the single most important reason why you didn't carry on with the TA?
What other things helped to make up your mind..?

REASONS FOR LEAVING	TOTAL* %	MOST IMPORTANT REASON %	OTHER FACTORS* %
JOB/MONEY	48	33	15
Demands of job	18	15	3
Change to shift work	16	12	4
Employer pressure	6	3	3
Not enough money	8	3	5
SOCIAL/TIME	46	23	23
Took too much time	16	7	9
Travelling time	5	2	3
Too many weekends	2	1	1
Affected family/social life	23	13	10
'INTERNAL' REASONS	47	27	20
Boring/lack of action	22	13	9
Discipline too strict	1	1	*
Discipline not strict enough	5	3	2
Didn't get skill training	6	3	3
Objections to higher ranks	5	2	3
Didn't get promotion	4	3	1
Equipment inadequate	1	1	*
Didn't like unit	3	1	2
PHYSICAL	10	7	3
Physical demands	5	3	2
Medical	5	4	1
OTHER	17	12	5

* Reasons for leaving were provided by respondents, therefore, these columns do not add up to 100%.

11. Pay and bounty (ie. annual payments contingent on meeting a minimum training commitment) were not perceived to be as important a contributory factor in wastage as might have been expected. Very few rated an increase in bounty as an improvement likely to improve retention, though its effect on recruitment and morale is likely to be more beneficial. A substantial increase in bounty or an alternative major financial inducement would be needed to improve retention significantly.

12. The quality of the training in the TA did not appear to be a major cause of wastage or dissatisfaction; the organisation of training events, however, particularly the frustration and inactivity caused by cancellations and delays was mentioned frequently.

13. The supply of modern equipment for training appears to be patchy but, on the whole, was not thought to be a major problem. Personal clothing and equipment allowances were criticised. Although it is thought unlikely to be an important factor in wastage, the cumulative effect of such minor irritants should not be underestimated.

14. The quality of both TA and Regular Officers and non-commissioned officers was considered to be reasonably high. The crucial importance of energy and imagination in the contribution of Regular training staff was stressed.

15. For those receiving supplementary benefit (ie. a UK welfare payment), and declaring it to the Department of Health and Social Security (DHSS), TA pay can result in benefit (and linked payments such as rent allowance) being reduced or eliminated, causing considerable resentment. Ignorance and rumour can make matters worse. More effective education and counselling for the unemployed TA recruit is recommended. A recent increase in the concession or "disregard" allowed by the DHSS from £4 to £8 per week may ease the sense or unfairness felt by those affected. The effects of this increase can be monitored and, if a problem continues to exist, the MOD will no doubt resume its efforts to persuade the DHSS to take a more sympathetic view.

16. Many leavers do not make a positive decision to leave. A temporary absence may be extended indefinitely through embarrassment, inertia etc. A more personal sympathetic follow-up of non-attenders (visits, telephone calls etc) by TA staff may reduce unnecessary wastage of this sort or, where this is unsuccessful, would improve the unit's knowledge of the reasons for wastage.

17. In the survey of "leavers" those who had stayed the longest tended to be older and more likely to be employed, married and to have children. They are slightly less likely than short stayers to have left school with academic qualifications. They are also more likely to have been members of uniformed youth organisations such as scouts, cadets etc. and to have had relatives in the Regular Army or the TA. This profile is of interest but of limited value in selection since the TA is unlikely to be in a position to reject suitable applicants on the basis of such factors.

18. Respondents were presented with five possible changes to terms and conditions currently under consideration (see Table 2). Nearly half said that one or more of the six would have made them less likely to leave but only the proposal for flexible terms of service (ie. the opportunity to earn some bounty by fulfilling less than current minimum commitment to obligatory training and annual camp) were regarded as particularly persuasive. While recognising the drawbacks, on the evidence of this study, we supported moves to increase flexibility in appropriate cases.

TABLE 2

POSSIBLE CHANGES TO TA TERMS AND CONDITIONS CURRENTLY UNDER CONSIDERATION

1. Splitting the bounty to allow payment of at least £100 as a bonus for attending camp, payable at camp.

2. Raising the third year bounty from £400 to, say £500.

3. Raising the first and second year bounties much closer to the third year bounty, say £300.

4. Issuing Rail Cards (ie. allowing cheap rail travel) to members of the TA similar to Student Rail Cards ie. for a small annual payment.

5. More flexible terms of service for the trained soldier to allow him to stay in the TA and qualify for some bounty for a lower training commitment - perhaps, an 8-day camp and a total of 15 instead of 27 days.

19. This study examined the wastage problem at a single point in time only. We concluded that there is a strong case for instituting a regular survey of currently serving TA soldiers and/or leavers in order to provide an up-to-date assessment of prevailing attitudes (and trends in attitudes) within the TA. This can in turn be compared to wastage rates, changes in terms and conditions and to external social and economic factors to provide more of the information needed for effective policy making.

20. Generally, the efforts made by central and local TA management to maintain the high quality of its internal operations and its relationships with the outside world were evident. Two suggestions for bringing the less effective units up to the quality of the best were:

a. Examine communications within the TA with a view to improving the spread of ideas and best practice.

b. In the overall evaluation of each unit's activities, which currently focusses on training effort and adminstration, an assessment of factors such as liaison with local employers, induction processes, and local efforts to reduce wastage should be included.

21. Local efforts to liaise with employers should be reinforced by a central initiative to educate, involve and organise employers.

SUMMARY

The study led to the following conclusions:

a. No dramatic improvement in the turnover and wastage rates from the present levels can be expected.

b. There is no single cause of wastage; job and social pressures, perceived characteristics of the TA, and a number of other factors all play a part.

c. No single measure will significantly affect wastage but a number of actions will improve the situation. These include greater flexibility in training commitments, the follow-up of non attenders, avoidance of inactivity during training events, counselling for the unemployed on DHSS benefit entitlements, and further attempts to improve the DHSS treatment of unemployed TA soldiers.

22. In addition to the particular measures covered in the study a number of general recommendations were made in the areas of communications within the TA, compilaton of statistical data, evaluation of units' activities, inplementation of a regular survey of attitudes, recruitment and selection, and the involvement and education of employers.

REFERENCES

Blyth D M. Study of Soldier Wastage in the Territorial Army. Report of the Army Personnel Research Establishment. November 1984.

Blyth D M and Dennison D. Retention in the UK Territorial Army. MTA Conference Paper, 1984.

Doering Z D and Grissmer D W. What we know and how we know it: A selection research and methods for studying active and reserve attrition/retention in the US armed forces. Second Symposium on motivation and morale in NATO Forces - contributions to the enhancement of the quantity and quality of military personnel, 1984.

Payne G D. Retention in the Army Reserve - A Review. Publication of the Australian Army Psychological Corps, 1981.

Shapland P C. Report of the Committee on the Study of Wastage. Internal UK Ministry of Defence Report, 1978.

Sub-Committee on Armed Forces of the Standing Committee on External Affairs and National Defence, House of Commons. (1981, December 18). Action for reserves. Hull. Quebec KIA 059: Canadian Government Publishing Centre. (Known as the Harquail Report.)

DIMINISHING RETURNS FROM AN ASSESSMENT CENTRE

by GAIL R HARDY

Department of Senior Psychologist(Naval), (SP(N)), Ministry of Defence UK.[1]

HISTORY

Royal Marine training has for many years had a high wastage rate, around 40%, but by the late Nineteen Seventies wastage was around 50%, and the Corps was consistently under-manned. In-house investigations had indicated ways in which the training regime and environment could be improved, but the climate of opinion amongst the trainers was that selection was the main problem. They were not convinced that the relatively brief selection procedure conducted in Careers Offices throughout the country was adequate to establish that the applicant had the physical fitness and motivation required to start training. As is so often the case, the trainers were inclined to feel that they could do a better job of selection than the Careers Staff.

In view of the urgency of the problem it was agreed to introduce a Potential Recruits Course (PRC) in October 1981, which combined some of the qualities of an Assessment Centre and a Realistic Job Preview. Each course lasts two and a half days and includes timed gymnasium tests, a timed run, an interview with a Personnel Selection Officer (PSO), psychometric tests, and other informative and recreational activities. Only applicants who have already satisfied the Careers Staff as to their suitability are forwarded to the PRC.

The aims of the PRC may be summarised as follows:

a. to determine whether the applicant is sufficiently physically fit to commence Royal Marine training, and has the necessary determination to overcome difficulties.

and b. to enable applicants to assess the type of life and training offered so that those who are not suitable may self-select out.

Additionally it was hoped that the experience would reduce the 'culture-shock' likely to assail new recruits, and that the trainers' morale would be improved by their participation in the selection process.

The final assessment and selection decision is based on marks awarded by three elements of the PRC team viz., a) the Physical Training

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of Senior Psychologist (Naval), or the Ministry of Defence.

Instructor (PTI) b) the PSO and c) the corporals looking after the course, who assess such factors as humour, cooperation etc.

It is worth noting that 64 recruit training places were lost in order to introduce the PRC, but the sacrifice was thought justified in the hope of reducing the excessive wastage rate. 64 places represents over 10% of recent annual recruit intakes.

EFFECTS OF THE PRC ON WASTAGE

PRCs were held twice weekly during training terms from 1st October 1981, and the first successful course members started training in late 1981/early 1982. Recruit training officially lasts six months, but back-trooping for various reasons can extend it to twelve months. Weather extremes during the year can also influence training results, so it was not until 1983 that a fair assessment of the PRC was possible. Statistical returns for recruits entering during calendar year 1982 showed a training wastage of about 25% - half the figure that had caused so much concern - and it would be fair to say that joy was unconfined.

The folly of unconfined joy became apparent when this author began to prepare a paper on the PRC (Hardy, 1984) to present to the 26th Annual Conference of the Military Testing Association. The first doubt related to the observation that training wastage had been declining steadily for two years before PRC-selected recruits entered training. The second was that the wastage graph for more recent recruits seemed to show a worrying upward trend.

Figure 1 demonstrates all too clearly that the second worry was justified.

As the PRC was introduced in October 1981, approximately one quarter of those entered in the year ending March 1982 had been through the procedure, as against approximately three quarters of those entered in the year ending September 1982. Wastage was down to 36% for the last year shown prior to PRC entrants, and 'bottomed out' for the year ending September 1982 at 24%. Therefore the maximum saving that could conceivably be attributed to the PRC is 12%, and that only if one chose to ignore the existing downward trend in wastage. As there were only 374 entrants in the year ending September 1982, a 12% saving represents approximately 45 recruits; rather less than the 64 training places that were sacrificed to accommodate the PRC.

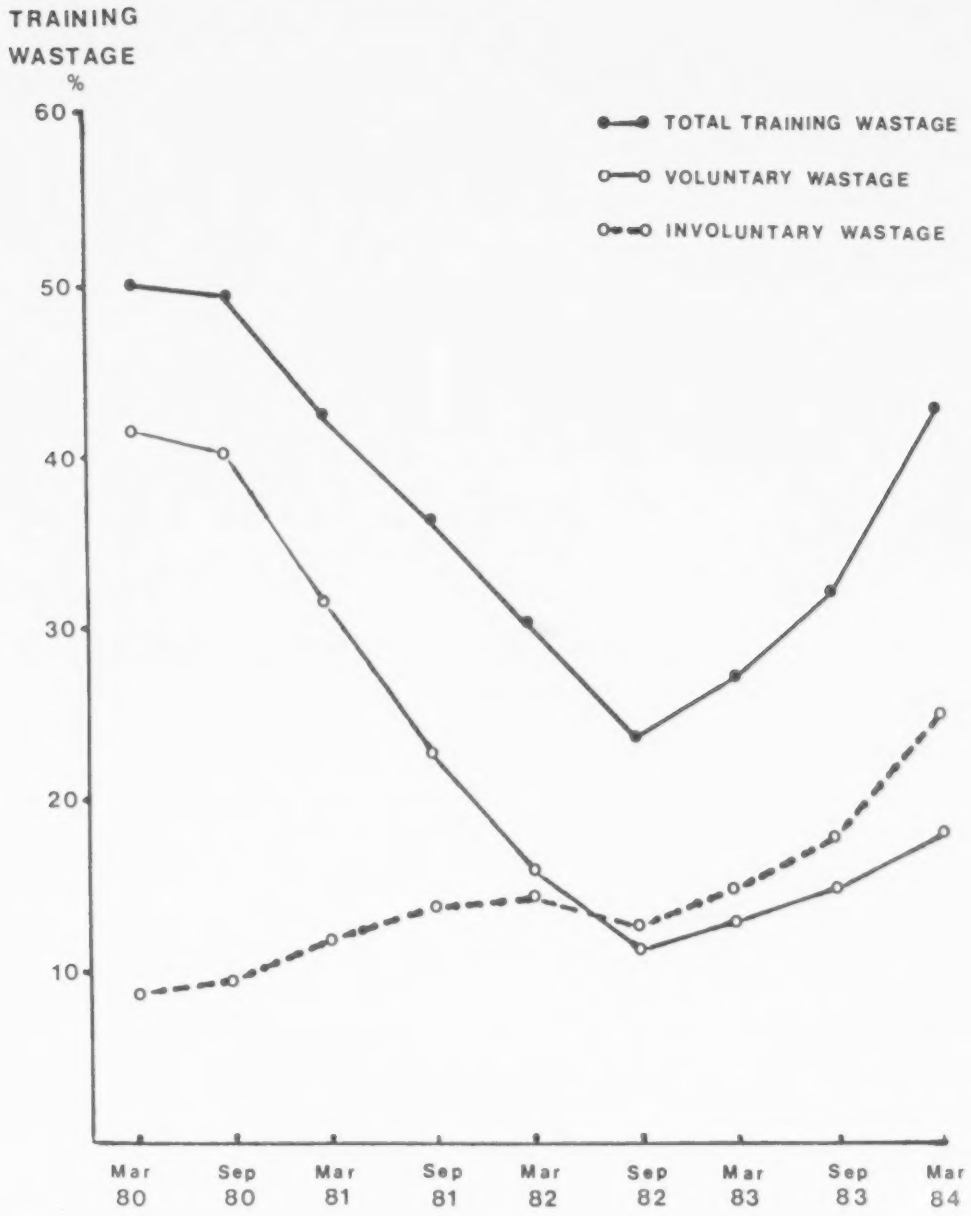
Of greater concern is the clear upward trend in wastage since the low for 1982 Entrants. For the year ending March 1984 wastage was around 42%, 6% higher than it was before the PRC was introduced.

A conspicuous feature of Figure 1 is the change in proportions between voluntary and involuntary wastage. In 1980 voluntary wastage was four

Figure 1

Training Wastage Rate Plotted Against Date of Entry

(Moving Annual Totals)



(Mean N = 887)

times as frequent as involuntary wastage, but recent figures show that involuntary wastage is now in the majority.

THE SEARCH FOR AN EXPLANATION

It has been argued that fear of unemployment was responsible for the original decline in voluntary wastage. As the unemployment of young people became more widespread, so the social stigma attached to that condition decreased, with the effect that unhappy recruits were again willing to leave. This is very probably true, but it does not account for the fact that the main increase in wastage is involuntary wastage, that is the trainers rejecting the recuit rather than the recruit opting out.

Early discussions of the problem with responsible authorities produced the comment that the quality of the recruits had declined, and this led to an examination of all the measurable and testable indicators of 'quality' held on the SP(N) data bank. These included psychometric tests scores, educational qualifications, age on entry, and even an assessment of 'Personal Qualities' made by the Careers staff. All these measures remained remarkably stable throughout the period in question, if anything showing a small improvement in some cases. Of course this does not disprove the 'quality' argument, but it makes it more difficult to sustain.

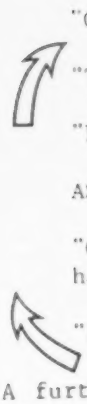
The early decline in voluntary wastage coincided with a marked reduction in the size of the recruit intake, and this led to some promising ideas that strain on training facilities and resources might contribute to voluntary wastage. Fleeting consideration was even given to a perverse type of equity theory, that is, whatever the size of the intake, whatever the selection method and whatever the quality of the recruit, the trainers would always produce the same number of trained Marines. Happily for those of us still clinging to the idea of an ordered world, the 'size of intake' theory has collapsed with more recent figures. Whereas an intake of 374 in year ending September 1982 had a wastage rate of 24%, an intake of 372 for the year ending March 1984 had a wastage rate of 42%.

When wastage trends were discussed with the Training Staff, PSOs and PRC personnel at the training establishment, there was a widely-held view that although training objectives had remained the same, trainers were informally raising their own standards to 'match' the improved capabilities of PRC-selected recruits. It is not in a keen young PTI's nature to allow a troop to lounge around if they have completed the assault course in half the allotted time. However, any unplanned raising of standards will mean that recruits who should have coped moderately well will have difficulty with training, and those who might have enjoyed training may become demoralised and opt to leave - or at least to stop trying.

The syndrome of diminishing returns - even from a successful assessment centre - is illustrated at Figure 2.

Figure 2

Diminishing Returns from an Assesment Centre


"Oh dear, half my trainees have failed".
"The recruits aren't up to standard".
"Let's only accept super-fit, motivated recruits".
ASSESSMENT CENTRE
"Gosh, these super-fit, motivated recruits are up to standard in half the time".
"Let's get them to do twice as much".

A further complication is the impact of the wastage rate on the PRC team. Members of the team are in informal daily contact with the trainers, and it is improbable that the trainers are restrained in their comments on 'poor' selections. Thus the PRC team are under continuous pressure to maintain or raise their selection standards. The PRC pass rate is relatively stable, although it does show a slight downward trend. However, the current pass rate of 55% is low in view of the fact that all course members have been tested and interviewed and found suitable at a Careers Office before being forwarded to the PRC. It seems likely that the standards being set by the PRC, particularly for physical fitness, are more relevant for Marines in training than for civilians wishing to start training. Anecdotal evidence suggests that an excess of zeal during some courses is responsible for several potentially good recruits withdrawing their applications. Both these factors reduce the potential recruit intake, and add to the problems of reaching complement.

HAS THE PRC FAILED?

It is impossible to know what the wastage rates would now be without the PRC - possibly even higher. However, on present evidence the PRC has not achieved what it set out to do. Wastage was falling before the PRC came into operation, and continued to fall as PRC selections started training. By 1983 wastage was climbing again and it is now higher than it was prior to the PRC.

It would be rash to attribute any of these effects solely to the influence of the PRC (or the consequent effect on the trainers). Unemployment, and changing social attitudes to unemployment, almost certainly have some, unquantifiable, effect. Similiarly, it is undeniable that when intake numbers were very high the relative

shortage of irons, drying facilities, telephones, etc for the numbers of users placed additional strain on the recruits. Implementation of some of the recommendations of earlier investigations very probably contributed to the initial fall in wastage. Nonetheless, this author does not believe that the PRC is justifying its existence on present performance.

It should not be thought that the PRC team, the training establishment, or their directing authorities are complacent observers of wastage trends. A detailed report on the problem has been prepared, and a thorough investigation and over-haul of the PRC is under way.

SOME LESSONS DRAWN FROM THE PRC EXPERIENCE

It is important to remember that although the PRC has many of the characteristics of an Assessment Centre, it is very much an in-house affair, only the PSO being professionally trained in selection. This is not to suggest that the other team members are amateurs, but rather that their professionalism lies in other spheres. However, this is probably true of many of the new crop of Assessment Centres, and the following observations may apply to them equally well.

a. An assessment team that is not working towards a definite quota of passes will tend to raise its' standards rather than accept 'training risks', particularly if it consists mainly of trainers. It is more socially acceptable to be seen to have high standards than low ones, and feed-back tends to concentrate on disasters.

b. An unsupervised assessment team may not set up any scheme to monitor and validate its' own performance, nor even be aware of the need to do so.

c. An over-rigorous assessment centre program may discourage potentially good recruits and cause them to withdraw their applications.

d. An Assessment Centre needs to know the quality of recruit it is required to produce. If the 'customer' informally raises his requirements to match the improved calibre of the recruits, the Assessment Centre is doomed to apparent failure.

REFERENCE

Hardy, G.R. An Assessment Centre and Realistic Job Preview for Selecting Men for the Royal Marines. **Proceedings of the 26th Annual Conference of the Military Testing Association** 1984. 1049-1054.

INCREASING BASIC SKILL LEVELS OF MODERATE RISK OFFICER CANDIDATES PRIOR TO COMMISSIONING IN HISTORICALLY BLACK COLLEGES

By
William D. Sprenger
Office of the Deputy Chief of Staff Army ROTC

It has been a source of some concern by authorities within the Army ROTC Headquarters that the basic skills of reading comprehension, writing, and computational ability were not as high as desired. Officer failures may take place at the initial training in an officer's branch which is called the officer Basic Course (OBC). It is here that young second lieutenants from the 16 branch schools across the United States are brought in order to learn the academics of their branch. From a total of 8,500 officers produced each year, the failure rate is about 1.4 percent. This may or may not be considered high, but to some it represents an unacceptable loss in manpower and resources. In 1984 there were 104 failures at the OBC and it was determined that 73 percent were caused by academic failures attributed to low ability to process information and/or poor writing skills.

Due to the complexity and range of quality of institutions of higher education in the United States, the guarantee of appropriate developed abilities cannot be assured--at least for ROTC requirements. Therefore, in 1983, Army ROTC felt the need to institute a qualifying examination titled the Officer Selection Battery (OSB), but this examination was subsequently disallowed primarily because of the adverse impact that it would have on minorities. The expected number of false negatives was deemed unacceptable. Partly as a result of this decision not to permit the OSB to be used as the sole screening vehicle into the advanced phase, an achievement testing program was initiated during the school year (SY) 1983-1984 that was to test the entire population of cadets (which at that time was about 72,000) in reading comprehension, English expression and mathematics. It is hoped that all of this will be completed and the new cognitive screens in place for the school year 1986-1987.

A concomitant benefit is that the testing program immediately places ROTC in the "value-added" business which is probably the most discussed program in higher education, although the techniques and its usage have been around for a considerable time.

Besides their appeal to the American psyche, value-added programs seem to be very worthwhile for certain institutions with certain types of students. In effect, ROTC will have approximately 1,000 "soft value-added" programs and 21 "hard value-added" programs. The qualifier "soft" is added since few contingencies are placed on the Professor of Military Science (PMS) to ensure completion of the program at this time, and since no criterion score is required for contracting or

commissioning (either in HBC or non-HBC) other than the PMS recommendation and a grade point average of 2.0.

Yoked to the achievement testing program is the main topic of this paper, which is the Enhanced Skill Training: ROTC Cadet Skills Development Program (EST). From 1916 to 1976, 21 Historically Black Colleges (HBC) were established as host institutions with an additional 53 formed as non-host schools. Only those enrolled at host colleges are eligible to participate in the skills program. These 21 schools account for 59 percent of the black officer production with the extension centers and cross-enrolled producing the remaining 41 percent (HBC only). Only those at HBC and enrolled in the EST courses have what could be a contingency in order to obtain a commission, that is, they must complete the courses in order to be commissioned, although in some cases doing so extends the college career by a semester, and in some cases, two semesters. The EST program is curious in that the federal government is involved in cooperating with 19 schools (two institutions elected not to participate during the SY 84-85) to produce higher basic skill levels. So far, the cooperation of the HBC has been nothing short of admirable even though both the schools and the Office of the Deputy Chief of Staff for ROTC had to work through the complexities of government contracting.

Only those cadets enrolled at the 21 host HBC colleges are eligible to participate in the skills program. These 21 schools account for 59 percent of the black officer production with the extension centers and cross-enrolled producing the remaining 41 percent (HBC only).

Over the years the HBC have produced exemplary officers, but due to several rather recent social-political factors, the percent of black OBC failures has increased and is considered unacceptably high. Of the total number of OBC failures, black officers (HBC and non-HBC) failures account for 55 percent with HBC contributing at the rate of 30 percent. The situation is becoming intensified because the Army intends to meet objective force requirements in 1990 and beyond, and to do so Army ROTC will be required to produce a steady state of 12,110 officers each year. Of this number, approximately 2,500 black officers will be required to mirror in leadership positions the enlisted force composition. To meet the challenge of producing black officers in the numbers needed and to avoid the adverse impact that a single mental screen would provide (such as the OSB) the EST program was launched in September 1985.

The essential features of the program are:

1. It provides additional classroom instruction funded by the Army that is designed to correct weaknesses in communicative skills, mathematics, reading, and cognitive skills.

2. Instruction complements rather than duplicates institutional subjects.

3. Instruction meets the professional education requirements for Army writing.

4. It requires individuals to be tested prior to entry in order to determine deficiencies and again at the end to measure progress made (in other words, a value-added program).

5. Instruction will be provided the enrolled cadet from the beginning of the freshman year up to the time of commissioning.

6. Instruction augments, but is not conducted in lieu of military science training.

7. A finite amount of funding is available through the contracting procedure which is based on ROTC enrollment figures. Cadets with the highest scores are to be selected for the enhanced skill training.

Each HBC will enroll contracted and non-contracted cadets if the cadet scores below the 6th stanine on the Stanford Mathematics Test, below the 6th stanine on the Missouri College English Test and below the 7th stanine on the Nelson-Denny Reading Test. A student may test out or exit the courses if he/she scores at or above the stanine mentioned above. For contracting purposes the school is considered to have met the terms of the contract if a student scores at the 5th stanine in mathematics, the 4th stanine in English, and the 5th stanine in reading.

Table 1 shows the mean percentile and raw scores on the achievement tests for the four ROTC regions and for males and females. As can be seen from the table, the Region means are ordered consistently across all tests, with Region IV being the highest followed by Regions II, I, and III. Region I includes those states up and down the East Coast, Region II is comprised of states in the Upper Midwest, Region III is located in the Central South and Southeast, and Region IV holds states in the Upper Great Plains and West.

These regional differences appear to reflect the distribution of HBC but one cannot assume that the HBC are the sole causal factor in mean score ranking. Female cadets scored higher than males on the English test, whereas males scored higher than females on the mathematics test. There were no overall differences between men and women.

Table 2 shows the means of HBC and non-HBC in each Region. Among the HBC, Region II means are the highest, followed by Regions I and III. Region II has the fewest HBC; Region IV has none. Among the non-HBC, Region IV scored the highest, but the means of Region I and II are not far below the Region IV means. Region II means are the lowest among the non-HBC schools.

The results of all tests indicate a sizable difference in the mean scores among racial groups. For example, on the English test, 93 percent of the white cadets attained the percentile score of 10 (raw score of 33) compared to 59 percent of the

black cadets and 67 percent of the Hispanic cadets. Of the total number, 15 percent tested would fall below this score (see the Total Column), but blacks and Hispanics would constitute a disproportionate number of those failing to achieve a raw score of 33. It is estimated that a raw score of between 45 and 50 is an indicator of at least average writing ability.

The HBC cadets have talent and brain power and one must not be led astray by average scores. It is the intention of Army ROTC to obtain from the HBC those cadets that have the highest test scores and those that show "reaction potential" toward EST. In any case, those cadets that are recommended for commissioning by the PMS will have scores in the range that will predict ability to compete in the Army system as beginning lieutenants as well as through the field grade ranks.

Table 1: Means of Percentile and Raw Scores by Region and Gender
(MS IV, SY 83-84)

		Mean Percentile	Mean Raw Score	No. of Cases	% of Total
English:	Total	40.2	52	7358	
	Region I	39.6	52	3097	42%
	Region II	42.1	53	1614	22
	Region III	35.5	50	1456	20
	Region IV	44.7	55	1191	16
	Males	38.9	52	6178	84
	Females	46.7	55	1180	16
Reading:	Total	43.0	126	7515	
	Region I	42.5	126	3262	43%
	Region II	45.5	128	1598	21
	Region III	36.7	119	1424	19
	Region IV	49.0	131	1231	16
	Males	43.3	126	6267	84
	Females	42.4	125	1196	16
Math:	Total	43.2	34	8215	
	Region I	41.9	34	3189	39%
	Region II	45.5	34.5	2032	25
	Region III	37.1	33	1664	20
	Region IV	50.5	35.5	1330	16
	Males	44.9	34.5	7001	85
	Females	33.2	32	1214	15

The percentile scores are based on the following comparison groups:
English: Four-year university and college freshmen (1964).
Reading: Four-year university and college freshmen through seniors (1981).
Math: College-preparatory high school seniors (1965).

Table 2: Means of Percentile and Raw Scores by Region and School Type (MS IV, SY 83-84)

		HBC			NON - HBC		
		Mean Percentile	Mean Raw Score	No. of Cases	Mean Percentile	Mean Raw Score	No. of Cases
English:	Region I	19.1	(40)	358	42.2	(53)	2739
	Region II	25.7	(45)	42	42.5	(54)	1572
	Region III	14.8	(37)	241	39.6	(52)	1215
	Region IV		NO HBC		44.7	(55)	1191
Reading:	Region I	13.1	(91)	328	45.8	(128)	2934
	Region II	25.5	(106)	42	46.0	(128)	1556
	Region III	10.5	(86)	237	41.9	(125)	1189
	Region IV		NO HBC		49.0	(131)	1231
Math:	Region I	18.1	(27)	382	45.1	(34.5)	2807
	Region II	19.1	(28)	34	45.9	(34.5)	1998
	Region III	16.1	(26)	252	40.8	(33.8)	1412
	Region IV		NO HBC		50.5	(35.5)	1330

The percentile scores are based on the following comparison groups:
English: Four-year university and college freshmen (1964).
Reading: Four-year university and college freshmen through seniors (1981).
Math: College-preparatory high school seniors (1965).

DIRECT MEASUREMENT OF ROTC CADETS' WRITING SKILLS

by

James P. Hanlon
Shippensburg University

We all agree that literacy is important. We believe it is especially important that our organizational leaders and managers are indisputably and demonstrably literate. But do we really know why literacy is important? Conventional explanations of the importance of literacy simply expand on the ideal of "an officer and a gentleman." That is, we rightly acknowledge how literacy enables a person to function, to perform his duty, to be an officer: That person must be able to transmit written messages clearly, concisely, and crisply: so also should he be able to receive them accurately and act upon them properly, with dispatch. No one, then, disputes the practical, functional role of literacy. On the other hand, we at least defer to, pay lip service to the idealistic role of literacy in officership. So literacy is likewise supposed to be an attribute of the gentleman--the person of sensibility, of humane tradition and values. Thus Breaker Morant, officer-hero of the fine Australian film bearing his name and depicting a tragic incident of the Boer War, was all the more a compelling protagonist because, in addition to being a fine soldier, he was as well an accomplished singer and poet: an officer and a gentleman.

But when the crunch comes, when we draw bottom lines, when we make hard budgetary and personnel decisions, we recognize the inadequacy of lip service and conventional reasoning. The functional rationale for literacy is all too self-evident: People who cannot deal competently with written texts sooner or later simply do not make it in the professional world we share. The idealistic rationale for literacy, on the other hand, is pretty much archaic cavalier and Victorian baggage--good enough for movies, I suppose.

Why then, is literacy important? To get at this question, we first should break literacy down into its two components: reading and writing. For writing is what I really want to talk about. Instead of viewing these activities as communications theorists do from a sending/receiving perspective, let's view them on a passive/active scale. Reading is essentially passive, even though more complex texts require higher-order interpretative, cognitive, and analytic skills. So reading may carry you from passive to active on my hypothetical scale. But writing begins with activity. And it may lead to true productivity, or pro-activity. A writer leads the reader through the text, whereas a reader follows a text. A fully literate person is a writer as well as a reader.

Only very recently has our educational establishment begun to seek out and deeply acknowledge the learning potential of writing. Put simply, writing is probably the best vehicle for learning and knowing in our educational arsenal. A good writer exhibits very high-level critical skills, reasoning ability, organizational skills, and --through awareness of audience--interpersonal skills. But, more importantly, through practicing the craft of writing, the student writer develops these skills. Perhaps writing is our only widely available means for developing these skills fully and collectively. Because we value these skills and abilities, because these skills and abilities contribute greatly to organizational effectiveness and dynamics, we seek and need potential career leaders who can and will write. A writer can reveal that he knows and what he really knows. In fact, through writing he discovers what he knows. So I believe this epistemological justification (writing as a way of knowing)

for writing provides the real, as opposed to the conventional, explanation for the high value we place on literacy.

Having then briefly advanced the fundamental argument for the importance of writing, allow me to share a few of my assumptions regarding the nature of writing, assumptions that are more or less germane to my proposal that the writing of officer candidates should be directly examined.

1) Writing is a technology: It is a very old technology, and a critical event in the European Renaissance was the invention of the printing press, which provided our first mass-produced consumer item, the book. As a technology, it can be learned on a skills-acquisition model, as opposed to a biological developmental model. Whereas we may be justifiably concerned if a young person does not develop listening and speaking capabilities by a certain early age, we should not make adverse judgments regarding a person's intelligence or learning abilities if that person has not acquired sophisticated writing skills by, say, the age of twenty. Nor should we blithely assume that, simply because that person has graduated from high school and perhaps done some college work, he is a poor learner of writing. He may well not have been instructed in writing, nor may he have been required to write in school, or anywhere else.

2) Writing is a discipline: Writing systems are wholly arbitrary and thoroughly conventional, nothing less than an elaborate social contract. Mastering a writing system requires incredible discipline--to say nothing of time. Further, the task of writing--of employing the system--is likewise exacting, frustrating, and time-consuming.

3) Writing individuates: No two persons will respond alike to a given writing situation. Whereas it is highly unlikely that a person would seek to copyright his response to the SATs (and, given the number of takers, he may have trouble proving his was original, for there is some statistical probability that other sheets are identical to his), that person could well seek sometime to copyright a love lyric, even though such lyrics abound in our culture.

4) Writing is not a rule-governed activity: We do not write, or learn to write, formulaically, by strict sequential prescriptions. To be sure, through automatization, we must acquire sufficient control of the surface features of the arbitrary writing system to focus on the content of our texts, rather than on the mechanics of producing them. A California throughway driver who has to think about where his brake pedal is and about which foot he wants to use on it is in BIG TROUBLE when, in a split-second of awareness, he knows he must stop his car instantly. So also is a writer in a different kind of trouble if he thinks predominantly about spelling and mechanics of punctuation as he writes. Unfortunately, this troubled writer may have been taught (should we say mistaught?) that spelling and mechanics are what writing is. In fact, writers develop incredibly varied, but seldom disclosed, strategies for composing. They learn these strategies through practice as well as through their experience as readers in learning how texts work.

5) Writing in the real world is a corporate activity: Almost all published, finished texts are produced through collaboration. Early drafts are shared and often benefit from early trusted-reader input. Many drafts are worked up and worked through in committee. Editors and advisors intrude and sometimes revise. Software programs mechanistically edit and silently correct predictable surface errors. Proofreaders check the final draft in scrupulous detail. Developing writers should

learn early how to buy into the corporate writing network, to use it to enhance and promote their authorship.

Since writing is, then, a corporate activity and since effective writing is essential to organizational welfare, I believe that there should be a larger measure of organizational concern for written performances within the organization. For example, many colleges and universities are developing programs of writing-across-the-curriculum. Such programs involve many faculty, rather than only English faculty, in teaching writing. Likewise, many colleges have developed writing centers and tutorial programs in writing to provide extra instruction in writing for those who want, need, or seek it. Graduate and professional schools are offering more and more writing courses for students seeking degrees in law, business, medicine, and even theology. So also are writing samples being used increasingly as part of selection criteria for admissions into both advanced undergraduate and graduate programs.

Officer training programs could as well use writing samples both as an instructional resource and as part of the selection process. I have recently proposed that the Army ROTC collect three writing samples from all cadets during their first two years in the program, prior to their contracting for the Advanced Course. Consider these justifications:

1) Writing samples provide direct evidence of writing ability: Other measurements provide only indirect evidence, even though such evidence may indicate very reliably how well the test-taker will perform in a writing course. In effect, measures of reading ability, of proofreading skills, and of common rhetorical ploys provide predictable measures of writing ability. But a series of writing samples provides a fully valid direct indication of writing ability.

2) Writing samples teach: Well-designed assignments show test-takers how to approach the task. Also, preliminary announcement of topics allows students time for the pre-writing tasks of talk and reflection. Likewise, people learn to write by writing, so each writing experience is a learning experience. Finally, the content of the writing could, through design of the assignment, relate directly to the ROTC Program of Instruction.

3) Collection of writing samples sends out the proper message: Consider this definition of motivation: People do what they have to do in order to do what they want to do. It follows, then, that a person who wants to be an officer will learn to write competently . . . if that person is certain he has to write competently. If that person is taking an indirect measure of writing ability, he will learn vocabulary words, proofreading rules for punctuation and capitalization, recognition of common grammatical errors, because these are things he has to know. But if that person is submitting a series of writing samples, then he will have to learn to write. In fact, you may see him skulking off to a writing center or seeking trusted-reader advice on a draft from a cadet who is recognized as a good writer. Likewise, through the simple expedient of allowing the cadet to submit as many samples, through time, as he wants to and of using only the final sequence of three as efforts of record, the cadet can then use the evaluation as a true learning experience and vehicle for improvement. He can, in effect, write his way toward commissioning.

Time and space do not allow me to elaborate upon the administration and evaluation of writing samples. Nonetheless, I will itemize some suggestions regarding procedures for handling and evaluating writing samples through the ROTC programs:

1) Each semester, PMSs should be provided with instructional packets for administering a writing sample. All first- and second-year cadets should submit samples each semester under carefully defined controlled circumstances.

2) PMSs should forward the sample to a contracted agency for holistic or primary trait evaluation on a 1-6 scale by two readers, and a third reader in the event of widely divergent first and second readings. The nationwide sample should be subjected to a common evaluation.

3) The contractor should report evaluations in a timely fashion to ROTC Headquarters, which would break out and distribute sets of evaluations region by region, detachment by detachment, cadet by cadet.

4) The contractor should store samples for a reasonable period of time, maintain accurate records of individual reader evaluations of each sample, and provide accurate records of reader reliability.

Without doubt, ROTC commitment to collecting and evaluating writing samples would be a considerable administrative and financial undertaking. But such an undertaking would provide indisputable evidence of the Army's commitment to quality in written communications at the entry level of commissioning. In the final analysis, this would be nothing less than a commitment to officer and organizational quality themselves.

819

PROCEDURES FOR REFINING WRITTEN MEASUREMENTS AT USAF OTS

SYDNEY SAKO
AND
LT COL WILLIAM J. SLAUGHTER

Officer Training School, USAF
OTS/MTC, Lackland AFB, Texas 78236

INTRODUCTION

Personnel at the Officer Training School, USAF, are always looking for ways to improve the product of the Officer Basic Military Precommissioning Course. During the past 25 years of school operation, with the production of over 80,000 graduates, one of the key factors of training has been the refinement of testing instruments. Among the several different types of measurements, the use of the consolidated written tests stands out as one of the major criteria in the graduation of students and selection of distinguished graduates. This paper presents some of the ways of refining written tests. It discusses test item writing, Test Review Board decision-making, item analysis, curriculum area achievement analysis, and statistical analysis of test scores. Information from these validational procedures was used in making decisions for conducting changes on specific items of the written tests. A total of over 200 test item revisions were made during 1983-1985. These revisions have resulted in improving significantly the test item bank by increasing the curricular validity of the items, lowering the standard error of measurement, and producing more reliable test scores.

POPULATION SAMPLE

The population sample used in this study consisted of over 5,000 students enrolled at OTS during the period of 1983-1985. Approximately 70 percent of the population have had no previous military service, and the remainder had served in enlisted rank, or active duty under the Airman Education and Commissioning Program. Their median age came close to 25, a drop from 27 observed during the years of 1978-1980. Almost 40 percent of them are married. All of them possess bachelor degrees, with approximately 5 percent having master or doctorate degrees.

MEASUREMENTS

At OTS we use a total of eight different types of measurements which include Consolidated Written Tests (CWTs), a Military Briefing, Graded Letters, a Communication Security Test, an Airman Performance Rating, a Flight Drill Performance Rating, Physical Fitness Tests, and two Officer Trainee Effectiveness Ratings. We have limited our study to the CWTs, since they are one of the most important criteria for graduation and for selection of distinguished graduates. The CWTs are designed to measure student achievement in the curriculum areas of Communicative Skills (CS), Leadership and Management (LM), Professional Knowledge (PK), and Defense Studies (DS). There are five different levels of CWTs used at OTS, with three alternate forms for each level. CWTs are administered at specific intervals in training; namely training days 15, 25, 34, 42, and 54. They are designed as criterion-referenced

tests, using a total of 235 test items that cover 67 criterion objectives of the course. (See Atch 1 for samples of criterion objective topics.) The minimum achievement standard for meeting the criterion objectives is set at 80 percent.

PROCEDURES AND RESULTS

The procedures used at OTS for test refinement include test item writing, Test Review Board decision-making, item analysis, curriculum area achievement analysis, statistical analysis, and pretest/posttest analysis. Pretest/posttest analysis is seldom used at OTS because of its high cost; therefore, it will not be discussed in this paper. The blueprint for test construction is the table of test specifications which is accomplished by a joint effort of testing specialists and subject matter experts. In order to maintain a high curricular validity, all test items are written by subject matter experts, also called Curriculum Area Managers (CAMs). Testing specialists make further refinement by using various statistical and other validation studies. The CAMs develop all course material and present lectures in their area of responsibility. They also write test items using well-established procedures which they have learned from attending the Test and Measurement Course on base and the 5-week Academic Instructor Course at the Air University. Once a CAM has developed a test item, it is submitted to a Test Item Review Committee for evaluation. This committee of other CAMs and division supervisory personnel critically review each test item, suggests improvements, and eventually approve every question before it can be included in a CWT.

During item analysis, we can tell how each item is functioning by looking at ease indexes and tabulation of alternate responses. As .80 is the minimum cut-off, any item with an ease index of .79 or below is checked for causes of low index reading. The cause may be due to ambiguity in item construction, inadequate emphasis of material, or insufficient assimilation or retention by students. An item analysis of over 5200 item administrations based on testing of 20 classes (Class 84-06 through 85-11) produced 271 items which fell below the ease index of .80. Flight commander (FC) comments and items below .80 were reviewed at the Test Review Board meetings attended by the chiefs of Curriculum Instruction Branch and Measurement Branch, CAMs, and FCs. The Test Review Board revised a total of 201 items.

The purpose of the curriculum area (CA) achievement analysis is to determine the amount of attainment of curriculum material in CS, LM, PK, and DS. This is accomplished by tallying the number of individual student failures in each flight, computing the percent failure rate by squadron, and then averaging the failure rate by class (see Atch 2). The goal is to achieve at least 90 percent in each area.

In CA achievement analysis categorized by class, the failure rate should not exceed 15 percent. If it is greater than 15 percent we must again determine whether the high failure rate is due to instruction, the test itself, or the curriculum material. In CA analysis by class, the failure rate was analyzed on 240 different CA administrations. Studies showed that the failure rates of 214 of them were less than 16 percent. The remaining 26 had failure rates of 16 percent or more distributed throughout the 21 classes (see Atch 3). The areas which showed the largest number of failures were DS with a total of 12, CS with 9, and LM with 5. There were none in PK which showed that all

students had passed the minimum requirements. This information is particularly useful to testing specialists, CAMs, and FCs. Testing specialists can pinpoint specific items with low ease indexes. CAMs can check their training material for any discrepancies, and FCs can emphasize those particular areas of difficulty during instructional hours.

In statistical analysis of test scores we obtain the differences of mean score (M), standard deviation (SD), reliability index (R), and standard error of measurement (SE) between the standardized and new tests. The norms of the standardized tests were obtained by averaging the data of operational tests administered to 20 classes. The new tests used in this study were administered to Class 85-11. Statistical results of the new tests were compared with those of the standardized tests, as shown below:

CWT 1	M	SD	R	SE
Standardized	90.1	5.6	.91	1.7
New	90.3	6.3	.90	2.0
Differences	.2	.7	.01	.3

CWT 2	M	SD	R	SE
Standardized	90.8	5.2	.90	1.7
New	91.7	5.6	.91	1.7
Differences	.9	.4	.01	0

CWT 3	M	SD	R	SE
Standardized	90.9	5.1	.90	1.6
New	89.9	5.8	.89	1.9
Differences	1.0	.7	.01	.3

CWT 4	M	SD	R	SE
Standardized	89.7	5.8	.89	1.9
New	87.5	6.5	.87	2.4
Differences	2.2	.7	.02	.5

CWT 5	M	SD	R	SE
Standardized	91.3	4.9	.91	1.5
New	90.5	4.3	.89	1.4
Differences	.8	.6	.02	.1

The differences between the standardized and the new tests were not significant; therefore, the new tests were put into operational status with only minor changes. By having the CAMs write the test items on their own material; by using the Test Review Board meetings to identify and take immediate corrective action on weak areas in testing, instruction, and curriculum materials; and by taking other curricular and test improvement actions when identified through item analysis, CA achievement analysis, and statistical analysis of test scores; we are continuously refining the testing and instructional programs at OTS in order to produce the finest young officers for the United States Air Force.

SAMPLES OF CRITERION OBJECTIVE TOPICS AT OTS

CS-2	EFFECTIVE WRITING
CS-3	COMMUNICATIVE PROCESS
CS-4	AF PUBLICATIONS
LM-1A	TIME MANANGEMENT & GOAL SETTING
LM-1E	STRESS MANAGEMENT
LM-2B	LEADERSHIP SURVEY
LM-2C	LEADERSHIP AUTHORITY
LM-2E	LEADERSHIP RESPONSIBILITY
LM-2F	LEADERSHIP CONCEPTS
LM-3C	MORAL OBLIGATIONS AND STANDARDS OF CONDUCT
LM-4E	GROUP DYNAMICS
LM-5A	MANAGEMENT IN THE AIR FORCE
LM-5B	MANAGEMENT FUNCTIONS & DUTIES
LM-5C	SUPERVISION
LM-5D	MOTIVATION THEORY
LM-5F	DELEGATION PROCESS
LM-6	COUNSELING
PK-1	AIR FORCE CUSTOMS AND COURTESIES
PK-2	AIR FORCE UNIFORM
PK-5	PROFESSIONAL RELATIONS
PK-6A	PAY, ALLOWANCES & LEAVE
PK-6B	MILITARY JUSTICE & DISCIPLINE
PK-8	PERFORMANCE RATINGS
PK-9	AIR FORCE DRUG/ALCOHOL PROGRAMS
DS-1	AIR FORCE HERITAGE
DS-2	MAN AND CONFLICT
DS-3	POWER IN CONFLICT
DS-4	DEPARTMENT OF DEFENSE
DS-5A	EMPLOYMENT OF AEROSPACE FORCES
DS-5B	STRATEGIC ISSUES OF THE 80s

Atch 1

766

853

SAMPLE

CURRICULUM AREA (CA) ACHIEVEMENT ANALYSIS

CLASS 85-06 CWT 4 TOTAL STUDENTS: 202 AVERAGE SCORES: 90.9

Part I: Number of Students Failed

SQ	FLT	STUDENTS	CS	LM	PK	DS	TOTAL CA FAILURES
1	10	19	4	0	0	1	5
1	12	18	5	0	3	2	10
1	14	20	6	1	1	2	10
TOTAL:		57	15	1	4	5	25
2	10	16	4	0	2	1	7
2	12	17	6	2	1	0	9
2	14	15	5	1	0	1	7
TOTAL:		48	15	3	3	2	23
3	10	18	4	0	0	1	5
3	12	17	4	1	0	0	5
3	14	15	2	0	2	2	6
TOTAL:		50	10	1	2	3	16
4	10	18	4	0	1	2	7
4	12	14	0	0	0	1	1
4	14	15	1	0	1	1	3
TOTAL:		47	5	0	2	4	11

Part II: Percent of CA Failures by Squadrons

	SQ 1	SQ 2	SQ 3	SQ 4	CLASS AVG
Communicative Skills	26.3%	31.2%	20.0%	10.6%	22.2%
Leadership and Management	1.7	6.2	2.0	.0	2.4
Professional Knowledge	7.0	6.2	4.0	4.2	5.4
Defense Studies	8.7	4.1	6.0	8.5	6.9

Atch 2

854

CURRICULUM AREA ACHIEVEMENT ANALYSIS
BY CLASS AND CWTs (PERCENT FAILURE RATE)

CWT	CURRICULUM AREA	84-06	84-07	84-08	84-09	84-10	84-11	84-12	84-13	84-14	84-15
1	CS 2,3	6	5	7	8	8	11	7	2	5	7
	LM 1A.2B.3B.4A.4E	5	(18)	6	5	11	7	6	2	8	13
	PK 1.2	4	9	9	8	4	3	5	11	6	10
	DS 1	13	(18)	10	8	5	13	8	9	4	12
2	CS 2.4	7	5	3	6	2	5	6	5	4	2
	LM 2C-H.5A.B	2	3	6	3	9	8	2	8	12	8
	PK 6	8	2	4	2	14	6	6	6	2	4
	DS 1.2	(18)	14	(29)	(27)	3	6	6	3	3	(18)
3	CS 2.5A	5	6	6	11	6	6	7	5	5	5
	LM 5C.5D.5F.5H	13	11	4	5	2	1	15	(17)	2	7
	PK 5A.8A1.9A	3	0	1	2	4	1	7	4	3	0
	DS 1C.3	10	14	8	6	(22)	8	5	(23)	9	15
4	CS 2	(20)	11	6	11	6	7	14	(19)	14	14
	LM 1E.3C	8	8	9	17	14	9	6	5	13	12
	PK 7.8	10	5	5	13	6	7	7	2	3	5
	DS 1D.4	10	5	6	8	12	11	7	9	14	7
5	CS 2	1	2	8	4	14	(18)	13	9	5	7
	LM 6	1	5	7	5	4	3	1	2	3	4
	PK 2C.5E.11.13.10	6	4	10	6	2	1	1	4	3	7
	DS 1E.5.6	1	4	4	4	10	15	4	6	9	5

CWT	CURRICULUM AREA	85-01	85-02	85-03	85-04	85-05	85-06	85-07	85-08	85-09	85-10	85-11
1	CS 2,3	8	9	9	4	0	4	13	6	2	12	11
	LM 1A,2B,4A,4E	(16)	2	11	8	0	6	2	14	12	14	(19)
	PK 1,2	7	3	5	6	2	4	7	2	6	5	9
	DS 1	5	5	6	1	0	1	1	1	0	4	3
2	CS 2,4	3	5	5	5	7	4	7	7	13	1	3
	LM 2C-H,5A,B	4	9	8	1	1	8	2	7	9	4	13
	PK 6	1	5	8	5	4	1	1	5	5	6	5
	DS 2	14	3	2	2	3	6	7	4	9	4	13
3	CS 2,5A	4	5	3	5	6	2	1	11	10	(20)	11
	LM 5C,D,F,H	11	2	2	1	4	4	9	4	5	3	0
	PK 5A,9A	1	2	3	2	2	1	4	2	7	3	11
	DS 3	(20)	7	8	5	4	5	5	7	11	10	13
4	CS 2	11	6	9	(23)	8	(22)	(18)	6	15	11	14
	LM 1E,3C	(16)	14	8	10	9	2	3	13	12	6	(16)
	PK 7,8	7	2	5	2	7	5	8	2	2	7	5
	DS 4	9	15	11	13	10	7	(17)	2	5	(16)	(22)
5	CS 2	13	7	8	9	8	11	15	11	(20)	7	10
	LM 6	1	2	5	1	2	2	1	4	1	1	4
	PK 5E,10,11,13	8	1	5	2	2	2	4	2	4	6	6
	DS 5	5	5	4	2	0	5	4	3	1	1	1

REFERENCES

1. Anastasi, Anne. 1982. Psychological Testing. 5th Ed. New York: MacMillan.

2. Bormuth, John R. 1970. On the Theory of Achievement Test Items. Chicago: Univ of Chicago Press.

3. Buros, O.K. ed. 1978. The Eighth Mental Measurements Yearbook, Vol 1 and 2. Highland Park, NJ. The Gryphon Press

4. _____ ed. 1974. Tests in Print II. HIghland Park, NJ. The Gryphon Press.

5. Cronbach, L.J. 1984. Essentials of Psychological Testing, 4th ed. New York: Harper and Row.

6. Edwards, Allen L. 1973. Statistical Methods. New York: Holt, Rinehart & Winston.

7. Green, J.A. 1975. Teacher-Made Tests. New York: Harper & Row.

8. Gronlund, Norman E. 1982. Constructing Achievement Tests. 3rd Ed. Englewood Cliffs, NJ, Prentice-Hall.

9. _____ 1985. Measurement and Evaluation in Teaching. New York: MacMillan.

10. Guilford, J.P. and Fruchter, B. 1977. Fundamental Statistics in Psychology and Education. 6th Ed. New York: McGraw-Hill.

11. Lyman, H.B. 1978. Test Scores and What They Mean. Englewood Cliffs,NJ: Prentice-Hall.

12. Mehrens, W.A. and Lehmann, I.J. 1984. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart & Winston.

13. Objectives & Tests. Vol III. AFP 50-58. Handbook for Designers of Instructional Systems, 15 Jul 1978. Department of the Air Force, Washington DC, USAF.

14. Randall, R.A. 1972. "Contrasting Norm-Referenced and Criterion-Referenced Measures." Paper presented at American Educational Research Organization,1972. Washington, DC.

15. Roid, G.H. and Haladyna, T.M. 1982. A Technology for Test-Item Writing. New York: Academic Press.

16. Sako, S. 1980. "Validation Studies of Written Achievement Tests used at USAF OTS." Proceedings of Military Testing Association 22nd Annual Conference.

17. Sako, S. 1983. OTS Measurements Development Handbook for Instructors. Lackland AFB, Texas.

18. Standards for Educational and Psychological Tests. 1985. Washington DC: American Psychological Association.

19. Student Measurements Administration Control and Special Individualized Assistance. 1984. Officer Training School Regulation 50-11. Lackland AFB,Texas.

20. Principles and Techniques of Instruction. 1984. Air Force Manual 50-62. Department of the Air Force.

856

Utility Estimation in Five Enlisted Occupations

Newell K. Eaton, Hilda Wing and Alan Lau

U.S. Army Research Institute for the Behavioral and Social Sciences[1]

In most organizations the decision to develop and implement selection and/or classification tests rests on the assumption that their costs will be outweighed by their benefits in terms of increased employee performance and tenure. The initial costs of testing programs have been increasing due to more stringent requirements for documentation of validities, test administration using computers, and the potential for legal challenges to test fairness. With the increasing costs of starting and maintaining testing programs, more attention is being paid to assessing their benefits. The purpose of this paper is to expand on methods used by several researchers in this area (Eaton, Wing, & Mitchell, 1985; Hunter & Schmidt, 1982).

Brogden (1949) and Cronbach and Gleser (1965) provided the first systematic descriptions of the utility of testing programs indexed in dollars. They linked performance levels to the dollar values estimated for those performance levels. Their formula for the gain in productivity, or utility (U$), obtained by using valid selection procedures includes (a) Ns, the number of individuals selected; (b) SD$, the standard deviation of performance, scaled in a utility metric such as dollars; and (c) the average performance expected on the criterion by the selected group as estimated from a valid predictor, given by Rxy Zx:

$$U\$ = Ns\ SD\$\ Rxy\ Zx$$

The formula was subsequently modified to account for testing costs. A more complete description of such formulations can be found in Cascio (1982), Cronbach and Gleser (1965), and Hunter and Schmidt (1982).

While the values of most of the variables on the right hand side of the Brogden-Cronbach-Gleser formulas are known, the estimation of SD$, the standard deviation of performance scaled in dollars, is problematic. One "SD$ Estimation Technique" is based on estimates of the dollar value to the organization of performance at the 50th percentile level, the 85th percentile level (one standard deviation above the mean), and, sometimes, the 15th percentile level (one standard deviation below the mean). The dollar difference between the 15% and 50% estimates, and the 50% and 85% estimates, provides an estimate of SD$ (Cascio & Silbey, 1979; Hunter & Schmidt, 1982; and Schmidt, Hunter, McKenzie, & Muldrow, 1979).

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

A second method is the 'Superior Equivalents Technique' proposed by Eaton et al.(1985). It is somewhat like the SD$ Estimation Technique. Instead of using estimates of the dollar value of 85th percentile performance, however, the technique uses estimates of the number (N85) of superior (85th percentile) performers who would be needed to produce the output of a fixed number (N50) of average (50th percentile) performers. This estimate, combined with an estimate of the dollar value (V50) of average performance, provides an estimate of SD$:

$$SD\$ = V50\ [(N50/N85) - 1].$$

Eaton et al. speculated that this method would be more appropriate in situations where the nature of the work is such that managers are more accustomed to considering the relative productivity of employees or crews than the relative costs of producing given levels of output.

A third estimation strategy has been proposed by Hunter and Schmidt (1982). In reviewing the results of a variety of studies, they note that SD$ typically falls between 40% and 70% of annual salary. This might be termed the 'Salary Percentage Technique.'

In their recent paper, Eaton et al. showed that the Superior Equivalents Technique provided more stable estimates of U.S. Army tank commanders' SD$ than did the SD$ Estimation Technique. They also noted that both these techniques provided substantially larger estimates of SD$ than did the Salary Percentage Technique. The purpose of this paper was to compare the results of the Superior Equivalents Technique with the SD$ Estimation Technique across five different U.S. Army enlisted military occupational specialties (MOS). This was intended to assess both the variability of SD$ values across the five MOS as well as the results with the two techniques. The paper was also intended to determine whether a 'short hand' estimation procedure could be developed for military occupations, such as the Salary Percentage Technique. Last, because the research was conducted with supervisors who were both noncommissioned officers (NCOs) and commissioned officers, it was possible to assess the impact of level of management on SD$ estimates.

METHOD

Instrument

A questionnaire based on earlier research (Bobko et al. 1983; Burke & Frederick, 1984; Eaton et al. 1985; Schmidt et al. 1979) was developed to measure the comparative worth to the Army of first-term soldiers operating at different performance levels. Separate forms were administered to supervisors in each of the five MOS studied. The first method asked supervisors to think about how much ten average soldiers (50th percentile) contributed to the Army. Supervisors then estimated how many superior (85th percentile) soldiers would be needed to do the same amount of work. The second method asked supervisors to first consider the worth of average and superior first tour soldiers to the Army. They were then asked to estimate how much an average (50th percentile) first-term soldier and a superior (85th percentile) soldier

are worth by considering such factors as salary, output, responsibility, and equipment. Dollar estimates of the yearly value to the Army of average and superior soldiers were then requested.

Subjects

Supervisory estimates were obtained from 270 NCOs and officers in five MOS across three different posts. The five MOS were infantrymen (11B), armor crewmen (19B), light wheel vehicle/power mechanics (63B), medical specialists (91B), and radio teletype operators (05C). Of the 270 supervisors, 226 (83 percent) were NCO and 29 (11 percent) were officers. The remainder did not provide rank information. Four supervisors (one percent) did not respond to the methods of estimating utility and their responses are not included in the analyses. Of the remaining 266, 13 did not provide useable estimates for the first method and (a different) eight did not provide useable estimates for the second method.

Other Data

To obtain the value of average performance for the Superior Equivalents Technique, as well as the data required for the Salary Percentage Technique, we used published pay and allowance tables. In 1985 the base pay for Army enlisted personnel with two years of service ranged from $9,000 to $10,000. Non-taxable allowances for such items as housing, post exchange, vacation, and travel benefits could amount to more than $6,000 for the typical married soldier with dependents. Our estimate of an equivalent civilian salary would be about $16,000 per year. This is consistent with Henderson's (1985) estimates for the compensation of a Private First Class living off post with dependents.

RESULTS

The results from the Superior Equivalents Techniques indicated that, across MOS, 5.20 superior first-tour soldiers performed as well as 10 average soldiers. Using $16,000 as the value of average performance (V50), 5.20 as the number of superior equivalents (N85), and 10 as the number of average soldiers (N50), the Superior Equivalents Technique yielded a SD$ estimate of $14,769. Of the 253 supervisors responding, 7% indicated 1 or 2 superior first-tour enlisted soldiers were equivalent to 10 average soldiers, 23% indicated 3 or 4, 51% indicated 5 or 6, 17% indicated 7 or 8, and 3% responded with 9 or 10. There was only a modest difference between estimates for the five MOS: the number of superior equivalents ranged from 4.90 to 5.58 with SD$ estimates from $12,881 to $16,720. The results by MOS are shown in Table 1. Full ANOVA results were computed, including as factors MOS and RANK of the supervisor providing the estimates. The differences by MOS did not reach statistical significance, nor did RANK, nor the MOS x RANK interaction.

The results from the SD$ Estimation Technique indicated that, across MOS, average soldiers were worth about $16,725 per year while superior soldiers were worth about $25,969. This yields an SD$ estimation of $9,244. Of the 258 supervisors responding, 11% provided SD$

Table 1: Estimated Number of Superior First Tour Soldiers Equaling 10 Average Soldiers and Computed SD$ by MOS

MOS	N	Number Superior	SD$
11B	48	5.54	$12,881
19E	60	5.40	$13,630
91B	36	4.89	$16,720
63B	67	5.15	$15,068
05C	42	4.90	$16,653
Totals	253	5.20	$14,769

estimates of less than $2,000, 14% between $2,001 and $4,000, 19% between $4,001 and $6,000, 12% between $6,001 and $8,000, 16% between $8,001 and $10,000, 15% between $10,001 and $16,000, and 12% over $16,000. These appear to be more variable than Superior Equivalents estimates. Larger, between MOS differences also were found with the SD$ estimation technique, ranging from about $6,254 to $11,150. The average values assigned average and superior soldiers, as well as SD$ estimates for the five MOS, are shown in Table 2.

Table 2: Dollar Estimates of Value to the Army of Average and Superior First Tour Soldiers by MOS

MOS	N	Average	Superior	SD$
11B	53	$19,226	$29,000	$ 9,774
19E	63	13,736	20,190	6,254
91B	38	18,000	27,132	9,132
63B	64	15,719	26,344	10,625
05C	40	18,200	29,350	11,150
Totals	258	$16,725	$25,969	$ 9,244

Full ANOVA results were computed on the SD$ estimates following the procedures outlined for the Superior Equivalents estimates. With the SD$ estimates, however, the effect of MOS was significant ($F = 4.23$, $df = 4,225$, $p < .01$). Duncan's multiple range tests indicated the SD$ estimates for first tour armor crewmen (19E) were lower than those for medics (91B) mechanics (63B) and radio telephone operators (05C). Neither the RANK nor MOS x RANK effects were significant.

Last, SD$ values obtained using both the Superior Equivalents and SD$ Estimation Techniques were compared to the estimated civilian equivalent salary and to base pay. Using $16,000 as the best estimate of estimated civilian equivalent salary and $9,500 as base pay, estimates of SD$ would be 58-92% of estimated civilian equivalent salary based on superior equivalents and SD$ estimates, respectively. Using only base pay as a salary basis, SD$ would be estimated at 97%-156%. The value of 125% of base pay may be chosen as an estimate of SD$. Assuming a value of $10,000 per year as base pay (for simplicity, rather than the $9,500 figure used in previous analyses), then SD$ = $12,500 and U$ can be estimated (Cascio, 1982, pp 220-226). Table 3 displays the estimated U$, per first tour soldier selected, as a function of the validity of the test and the proportion of applicants selected.

Table 3: Estimated U$ Per Selection as a Function of Test Validity and Proportion of Applicants Selected

		Test Validity					
		.1	.2	.3	.4	.5	.6
Proportion	.2	$1,750	$3,500	$5,250	$7,000	$8,750	$10,500
of	.4	1,200	2,400	3,600	4,800	6,000	7,200
Applicants	.6	813	1,625	2,438	3,250	4,063	4,875
Selected	.8	413	825	1,238	1,650	2,063	2,475

If 100 soldiers were selected from among 125 applicants, using a test with a validity of .3, the estimated utility would be 100 x $1,238 = $123,800 per year.

DISCUSSION

The first purpose of this research was to assess the SD$ of performance in five Army enlisted military occupational specialties using two methods. For both methods there were numerical differences in SD$ across the MOS, and they were ordered logically. The lowest SD$ values were obtained for team/crew occupations - infantryman and tank crewman - while the highest SD$ values were obtained for those who perform many duties as individuals - medics, mechanics, and radio/telephone operators. However, between MOS differences were statistically significant for SD$ values obtained for only one method, the SD$ Estimation Technique, and these differences were not clear cut.

The Superior Equivalents Technique, designed for use in military settings, did not provide reliable between-MOS differences. It did, however, yield estimates with considerably smaller levels of between-subjects dispersion. This is consistent with the results of the earlier Eaton et al. research. On balance, it would seem that both techniques provide SD$ estimates which yield a useful range in which the "real" SD$ probably falls. But neither is sufficiently precise at this time to provide between-MOS comparisons in which one can be confident.

Obtaining a ball-park estimate of SD$ may well be sufficient for most purposes. Seldom does one face a decision where the utilization of a selection or classification test rests on cost tradeoffs of plus or minus 10% or 20% of testing and start up costs. Rather, such programs are more typically initiated only if the potential payoff is several times the costs. As a consequence, estimating a reasonable range of SD$ values can be quite useful.

Fortunately, this and prior research (Eaton et al. 1985) show that such an estimate may be obtained using a variant of the Hunter & Schmidt (1982) Salary Percentage Technique. In the Eaton et al. work, SD$ was 89% of estimated civilian equivalent salary, and 178% of base pay. For the two methods compared in this research, results ranged from 58%-92% of civilian equivalent salary, and 97%-156% of base pay. Given this consistency it would seem that a rough estimate of SD$ for first-tour enlisted personnel is about 125% of base pay.

Such an estimate is likely to be quite conservative. Eaton et al. found SD$ values obtained with the two methods used in this research to be about half those obtained with yet a fourth method, the System Effectiveness Technique, designed to incorporate equipment, maintenance, and other support costs. Burke and Frederick (1984) and Schmidt et al. (1982) also obtained results suggesting the conservative nature of SD$ values obtained with the SD$ Estimation Technique. The use of such a rough estimate may well make a useful contribution to front end analyses designed to assess the potential utility of initiating research on, or implementation of, a selection and classification testing program. Table 3 provides figures which make such estimates relatively simple.

REFERENCES

Bobko, P., Karren R., & Parkington, J.J. (1983). The estimation of standard deviations in utility analyses: An empirical test. Journal of Applied Psychology, 68, 170-176.

Brogden, H.E. (1949). When testing pays off. Personnel Psychology. 2, 171-183.

Burke, M.J., & Frederick, J.T. (1984). Two modified procedures for estimating standard deviations in utility analyses. Journal of Applied Psychology, 69, 482-489.

Cascio, W.F. (1982). Costing human resources: The financial impact of behavior in organizations. Boston: Kent Publishing Co.

Cascio, W.F., & Silbey, V. (1979). Utility of the assessment center as a selection device. Journal of Applied Psychology, 64, 107-118.

Cronbach, L.J., & Gleser, G.C. (1965). Psychological tests and personnel decisions (2nd ed.), Urbana: University of Illinois Press.

Eaton, N.K., Wing, H., & Mitchell K.J. (1985). Alternate methods of estimating the dollar value of performance. Personnel Psychology. 38, 27-40.

Henderson, W.D. (1985). Cohesion: The human element in combat. Washington, D.C., National Defense University Press.

Hunter, J.E., & Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In Fleishman E.A., Dunnette M.D. (Eds.), Human performance and productivity: Volume I. Human capability assessment. Hillsdale, NJ,: Erlbaum.

Schmidt, F.L., Hunter, J.E., McKenzie, R., & Muldrow, T. (1979). The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 64, 609-626.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work force productivity. Personnel Psychology. 35, 333-347.

RELATION OF MENTAL AND EDUCATION LEVELS TO NAVY ENLISTED PERFORMANCE

Charles H. Cory[1]

Navy Personnel Research and Development Center

As part of a joint services effort, the Navy Personnel Research and Development Center has been investigating the validation of enlisted personnel selection standards against job performance as well as the training school achievement criteria presently used. Personnel-record based job outcome measures are one type of criterion being investigated. This paper summarizes the findings from two studies which measured the association of mental and education levels of Navy enlisted personnel to their performance on job outcome criteria.

The first study, conducted on the CY76 male enlisted cohort, investigated the relationship of education (ED) and mental level (ML) to 6 job outcome criteria--2 advancement, 2 attrition/retention, and 2 maladaptive behavior criteria. Relationships were examined within the three entry pay grades in the cohort: E-1, E-2, and E-3.

The approximately 71,000 male, non-prior service personnel accessioned during CY76 who were 4-year enlistees and had complete records on the variables of interest constituted the sample. The predictors were: (1) Education Category coded with three values: Non-high school Graduate=1, GED Certificate=2, and High School Diploma Graduate=3 and (2) mental level category(ML), coded with five values: 93-99th centile(ML 1)=1, 65-92nd centile(ML 2)=2, 49-64th centile(ML High-3)=3, 31-48th centile(ML Low-3)=4, and 10-30th centile (ML 4)=5.

The two advancement criteria were: (1) percent of time spent at E-4 or higher(PTE4+) and (2) achievement/non-achievement of E-4 during the 4-year enlistment(E4/Non-E4). The attrition/retention criteria were: (1) completion/non-completion of enlistment and (2) attrition/non-attrition within one year of accession. The maladaptive behavior criteria were: (1) number of demotions and (2) overall behavior record, a composite score formed from the number of unauthorized absences + twice the sum of the demotions and desertions.

Validities of ED and ML across Ratings

Figure 1 shows the validity coefficients of ED and ML for the two advancement criteria. The four columns for each predictor represent the validity coefficients for the subsamples accessioned as E-1s, E-2s, E-3s, and the total sample, respectively. Because the scaling of ED and ML are arbitrary and the focus of the research is to compare their magnitudes, the absolute values of the validity coefficients are shown. In other words, for this figure, the negative coefficients for ML are presented as positive values.

---

[1]The opinions expressed in this paper are those of the author, are not official, and do not necessarily reflect the views of the Navy Department.

The top part of the figure shows that both education and mental level had substantial effects on percent of time spent at E-4 or higher during the first enlistment. Validity coefficients for ED range from .21 to .11 and the validity of ED overall, without subgrouping by entry pay grade, was .24. Validity coefficients for mental level ranged from .10 to .25 and the overall coefficient was .27. The predictiveness of ED and ML for E4/Non-E4 was similar to but slightly lower than for percent of time at E4+.

For attrition/retention criteria (Figure 2) the predictiveness of ED was from two to three times that of ML, depending on the criterion and the entry pay grade group. For Completed the 4-year Enlistment, validity coefficients for ED ranged from .20 to .06 compared with a range for the ML coefficients of .01 to .03. For Attrited during the First Year, validity coefficients for ED ranged from .14 to .07, compared with a range of .03 to .06 for ML.

ED was also more predictive for Maladaptive Behavior criteria (Figure 3) than was ML. For Overall Behavior, validity coefficients of ED ranged from .18 to .08. Those for ML ranged from .01 to .02. Coefficients of ED for Number of Demotions ranged from .09 to .03, those for ML ranged from .01 to .02.

Validities of ED and Mental Ability Within Ratings

The predictiveness of education and mental ability variables for job outcome criteria also were studied within ratings. The sample selected for this research consisted of personnel in three representative ratings: Sonar Technician (Surface), Hospital Corpsman, and Machinist Mate. The samples were taken from a data base composed of longitudinal records which covered the 4-year enlistment outcomes of all first-term non-prior service male personnel who were accessioned during FYs 77, 78, and first quarter 79.

The data base contained complete rating and pay grade records of personnel during the 4-year period, or that part of it during which they served in their first enlistment, and a calculation of total basic pay, based on time in each pay grade. Samples were selected to include all 4-year non-prior service male enlistees who at some time in their four years had been either: (1) assigned to the rating, (2) assigned to formal school training for the rating or (3) promoted to E-4 or higher in the rating. Sample sizes were: 956 Sonar Technicians, 2,278 Hospital Corpsmen, and 3,177 Machinists Mates.

For this study the only criterion used was time at E4 or higher, the criterion which had been found to be the most promising in previous research. This was measured as Months at the Full Performance Level (MFPL), i.e., time spent at E-4 or higher.

Figures 4 and 5 show the predictiveness for MFPL of the selector composite which is used for classification into the rating and the three variables which define Navy requirements for enlistment. These are: AFQT score, Education level categorization (ED), and SCREEN score. SCREEN, an acronym standing for Success Chances of Recruits Entering the Navy, is a composite score, scaled from 48 to 96, which is based on AFQT, education, age and marital status. Personnel with scores of 70 or higher are eligible for Navy enlistment. SCREEN was developed to exclude from Navy enlistment personnel who are high attrition risks.

Figure 4 shows,for personnel accessioned as E-1s, the predictive validity of AFQT, ED, SCREEN, and the selector composites for the three ratings. Sonar Technician has the highest coefficients overall, and Hospital Corpsman and Machinist Mate follow in that order. Validity coefficients for STGs range from .10 to .27, with similar ranges for the other two ratings. All 12 validity coefficients are statistically significant. For the three ratings SCREEN was the best predictor of MFPL and ED was a better predictor than AFQT.

Figure 5 shows for personnel accessioned as E-3s, the same type of predictive information shown previously for personnel accessioned as E-1s. These coefficients are much lower than those for E-1s. Validity coefficients for Hospital Corpsman, the most predictable rating, range from .11 to .17 and those for Machinists Mate, the least predictable rating, range from .01 to .06. Only 7 of the 12 coefficients in the slide are statistically significant.

Comparisons of this type are not being presented for E-2s because E-2 accessions are much less common than E-1 and E-3 accessions. E-2s constitute only about 7% of enlisted accessions, compared with 74% accessioned as E-1s and 19% accessioned as E-3s. Therefore, the sample sizes for E-2s are much smaller than those for the other two levels, and the predictive relationships are more irregular.

Relative Utility of Personnel Accessioned as E-1s, E-2s, and E-3s

The same data base was also used for calculations to develop a dollar criterion of utility. Because the utility relationships were similar for the three ratings, only those for Machinist Mate are presented. Figure 6 shows the relative performance of personnel accessioned as E-1s, E-2s, and E-3s in terms of mean months at full performance level, mean total basic pay cost per month at FPL, and first term attrition rate. All four variables have been scaled in multiples of the lowest value. Thus, for the leftmost variable, the mean MFPL of E-1s is scaled as 1.0 that for E-2 is roughly 25% greater, etc.

Mean months at full performance level for E-3 accessions were more than twice those for E-1s and about 80% greater than those for E-2s. In contrast, the mean cost of E-3s per month at FPL was about 35% that of E-1s. The reasons for this large discrepancy in mean basic pay cost per month are shown in the next two categories. Mean total basic pay costs for E-3 accessions for the entire enlistment were about 4/5ths those of E-1s and E-2s. Since E-3s are paid more than E-1s and E-2s and they spend more time at E-4+, one reason for their surprisingly small mean total basic pay is shown in the next category to the right: E-3 accessions attrited at nearly twice the rates of E-1s and E-2s. Attrition rates of 28.0%, 24.9%, and 47.3% were recorded for E-1, E-2, and E-3 accessions, respectively.

Statistical Corrections to Improve Estimates of Validity

Although these analyses are preliminary, one gratifying result from them is that they show that selection and classification variables in the Navy are associated with important aspects of enlisted performance. Moreover, it is clear that the predictivenesses of ED and mental ability for job outcome criteria are even greater than is shown in the preceding figures. For instance, inspection of

the AFQT means and standard deviations for the samples used for the research showed that restrictions in range had occurred. Correction of the validity coefficients for these restrictions in range will produce a more accurate measure of the predictive relationships; it will also show them to be larger than the coefficients presented in Figures 1 through 5.

Also, because the shapes of the distributions of both ED and ML differ substantially from normality, the actual maximum validity coefficient is substantially less than the theoretical maximum, 1. The departure from normality was the most pronounced for the ED distribution: 19% Non-HS Grad, 7% GED, and 74% HS Grad--an asymetrical U-shaped distribution.

Estimates of the effects of these departures from normality on the effective maximum validity coefficients for the predictors in the CY76 cohort were made. For this step a method recommended by Carroll was used to calculate the effective maximum correlation coefficient, given the non-normal characteristics of the ED and ML distributions for two criteria: Percent of time spent at E4+ and Completion/Non-completion of Enlistment. The maximum coefficients for ML for PTE4+ and Completion/Noncompletion were .91 and .57, respectively, and for ED, .42 and .64.

This substantial shortfall of the maximum rs from the theoretical maximum, 1.00, indicates that both ED and ML are more strongly related to the performance criteria than the absolute magnitudes of their validity coefficients indicate. This is particularly true for ED-PTE4+. ED accounts for 33% of the explainable variance of PTE4+, rather than 6% of the variance as is indicated by the absolute magnitude of the computed coefficient.

Summary and Conclusions

These research findings support the following conclusions: (1) The education level, mental ability and SCREEN variables which are used for enlisted selection in the Navy are significantly predictive for important job outcome performance criteria, both within and across ratings. (2) ED and mental ability are about equally predictive for advancement criteria, but ED is much more predictive than mental ability for attrition/retention and maladaptive behavior criteria. (3) Selection and classification variables are more predictive of job outcome variables for E-1 accessions than for E-3 accessions. (4) Adjustment of the data to compensate for the distortions caused by departures from normality of the distributions of the predictors and restriction in range can be expected to show that the true predictiveness of these variables for job outcome criteria is even greater than the substantial coefficients found in the present research. (5) The practice of accessioning qualified personnel at E-2 and E-3 pay grades is a very cost effective way of providing to the Navy enlisted personnel who can perform effectively at the full performance level.

This report describes preliminary findings from research designed to develop a methodology for calculating a composite job outcome measure which expresses an individual's overall value to the Navy. The research will continue during FY86 and beyond.

VALIDITY COEFFICIENTS OF ED AND ML FOR ADVANCEMENT CRITERIA

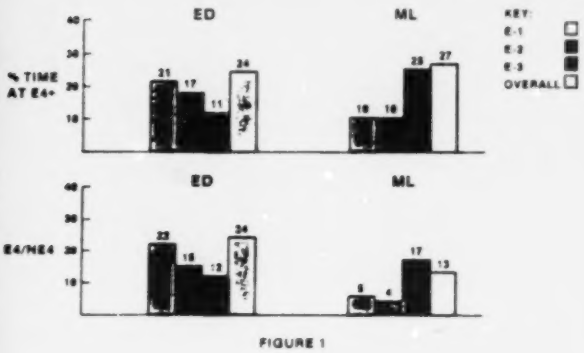


FIGURE 1

VALIDITY COEFFICIENTS OF ED AND ML FOR ATTRITION/RETENTION CRITERIA

FIGURE 2

VALIDITY COEFFICIENTS OF ED AND ML FOR MALADAPTIVE BEHAVIOR CRITERIA

FIGURE 3

PREDICTIVE VALIDITY OF AFQT, LEVEL OF EDUCATION, SCREEN, AND SELECTOR COMPOSITE FOR MFPL (FOR E-1 ACCESSIONS)

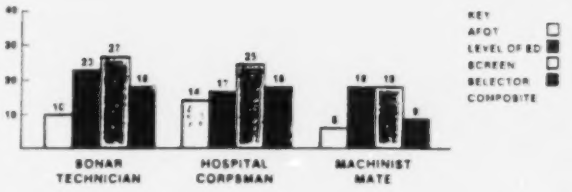


FIGURE 4

PREDICTIVE VALIDITY OF AFQT, LEVEL OF EDUCATION, SCREEN, AND SELECTOR COMPOSITE FOR MFPL (FOR E-3 ACCESSIONS)

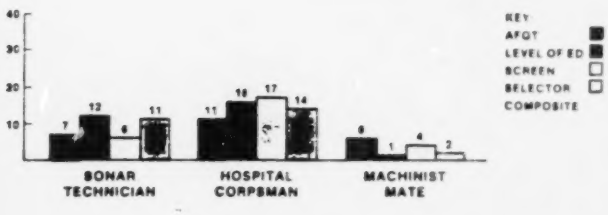


FIGURE 5

ATTRITION AND ADVANCEMENT CHARACTERISTICS of E-1, E-2, and E-3 MACHINIST MATE ACCESSIONS

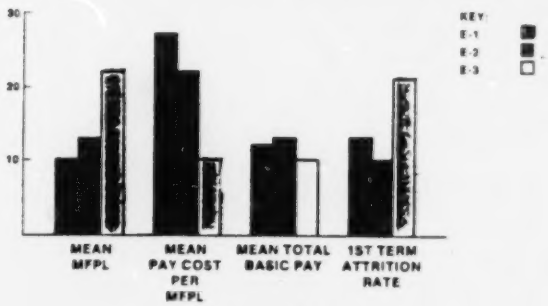


FIGURE 6

SELECTING CRITICAL TASKS FOR A RADIOMAN HANDS-ON PERFORMANCE TEST

Herbert George Baker, PhD
and
Gerald. J. Laabs, PhD

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Selecting an array of critical tasks is one of the most important steps in developing a hands-on job sample test. This paper describes the procedures used in critical task selection for the Radioman (RM) performance test and for surrogate instruments.

Background

In cooperation with the Joint Service Job Performance Measurement/Enlistment Standards Project, the Navy is exploring relationships between predictors and various measures of job performance. The main objective is to investigate measurement approaches that might be used to make personnel classification more performance-based. The outcomes of the research may allow a recruit's potential for successful on-the-job performance to be considered more fully in the personnel accessioning process, perhaps through the addition of an additional algorithm component in the automated classification and assignment system.

The observation of job performance would seem to be the ultimate basis upon which to validate tests used to screen military applicants. Thus, the strategy of the Joint-Service Project is to construct hands-on performance measures and investigate the use of these measures as criteria for predictor validation, with a major research focus on the development of job sample tests of technical proficiency.

The institution of a service-wide program to obtain performance criteria via job sample testing would be prohibitively expensive. Therefore, a concurrent research focus of the Joint-Service Project is the development of economical substitutes, or surrogate tests. Costs will be reduced by using hands-on performance tests as benchmark performance measures and developing substitute measures that are less expensive and easier to administer, such as paper-and-pencil or computer simulations of job samples.

Radioman Performance Test Development

The Radioman rating is one of the jobs selected by the Navy for performance test development because it is critical to mission success, has a large population -- including substantial numbers of women and ethnic minorities -- and is similar to radioman jobs in the other armed services. In accordance with the Joint-Service research strategy a job sample test is being designed and developed, along with a job sample simulation and a set of rating scales. This work is being performed under contract by the Human Resources Research Organization (HumRRO) and the Personnel Decisions Research Institute (PDRI).

868

Critical Task Selection

For the job sample test under construction to be valid, it must adequately represent the important tasks done by first-term radiomen. In fact, the validity of the entire test package -- the hands-on test, simulation, and rating scales -- depends on the method used to identify and select tasks to be tested. Critical task identification and selection is the first and perhaps most important step in the test development process.

The Navy and its contractors have completed this first step, gathering information on critical tasks from subject matter experts (SME) who are senior enlisted personnel in the RM rating. As an additional check on the data gathered from RMs working in the fleet, a Quality Control Review Panel (QCRP) composed of additional SMEs was formed. This was done by soliciting nominations for panel membership from headquarters, school, and operational Navy commands that have cognizance over the RM rating. This panel will provide expert review of the work at each step in the process, assuring the highest quality end-products. In addition, it will establish an audit trail, facilitating acceptance of the final test package by the commands whose representatives are on the panel.

Out of the plethora of tasks done by RMs, a manageable subset must be selected as candidates for the development of hands-on test items. These items will be the critical job tasks that can feasibly be measured in the hands-on mode, and in turn will guide the development of substitute, or surrogate, test items. To guide critical task identification and selection, we have adopted Guion's (1979) paradigm for reducing the job to a job sample (Figure 1). His four major steps include determining the: (1) job content universe; (2) job content domain; (3) test content universe; and (4) test content domain.

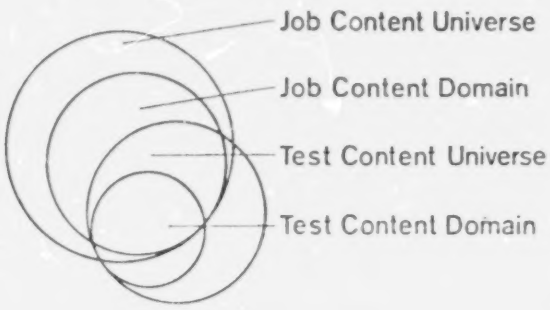


Figure 1. Venn diagrams relating job content to test content (adapted from Guion, 1979)

Job Content Universe

The RM job content universe was defined by a comprehensive job task analysis. Two sources of information were used: Data from the Navy Occupational Task Analysis Program (NOTAP), which is used to develop the RM occupational standards; and the Job Task Inventory (JTI), which is used in developing the entry level training (i.e., A-School) RM curriculum. The two task analyses overlap considerably and contain 500-600 different job task statements.

Job Content Domain

Next, the job content domain for first-term RMs was specified. The original list of task statements was refined and reduced through two workshops that included SMEs who were supervisors of first-term RMs. Both ship and shore installations were represented. Seven senior Navy enlisted personnel and one civilian attended the first workshop which was held in San Diego. Then, the shortened list was further refined at a second workshop held at Norfolk, in which seven senior Navy enlisted personnel participated.

To ensure that all task statements were written at the same level of generality, subtasks were accumulated into more comprehensive tasks, while broader, more general tasks were separated into their constituent parts. The Joint-Service research strategy addresses technical job proficiency only; that is, it specifically excludes purely military or routine administrative tasks, as well as other job performance considerations such as team tasks, and motivational, situational, or stress factors. The workshop panels deleted tasks falling into those categories, as well as tasks not within the normal purview of first-term RMs.

A list containing 125 task statements resulted. These statements were evaluated by the review panel, who gave approval to a final list that included 127 task statements. These tasks comprise the technical proficiency domain of first-term RMs.

Test Content Universe

The next step in the procedure was to determine the test content universe. Put simply, this universe potentially embraces all those tasks that could possibly be included in a hands-on test, plus those elements necessitated by the testing situation. The testing situation includes such things as those conditions that are imposed to achieve relatively standardized testing and the procedures used to observe and record responses. For the RM rating, the test content universe is essentially the same as the job content domain, except for the addition of the required testing conditions, and a way must be found to reduce the size of the test content universe to manageable proportions. That is, some priorities must be established to guide task selection.

One of the typical methods to set task selection priorities is to gather information on the criticality of performing each task correctly. Consequently, the Navy launched an extensive survey effort to gather judgments from personnel in the rating, for use in setting task selection priorities. The 127 task statements were incorporated in a questionnaire, the Radioman (RM) Survey Form, designed to isolate the critical first-term RM job tasks.

The review panel addressed four proposed judgment scales:

1. frequency of task performance by first-termers,
2. difficulty or complicatedness of the task,
3. importance of the task to mission accomplishment, and
4. frequency of errors on performing this task by first-termers.

As a result of the review, the scales were changed to reflect absolute rather than relative judgments, with the review panel providing the scale values or anchors. The questionnaire was developed in two forms: one for first-term RM personnel (defined as RMs with four or less years of active duty), and one for supervisors (those RMs with between four and ten years of active duty). First-term RMs judged the frequency and difficulty of the tasks; supervisors judged the task importance and error frequency. The RM Survey was pilot tested on board a ship, using a small sample of RMs and supervisors.

The review panel also assisted in developing the sampling plan for distribution of the RM Survey Forms. Panel consensus was that the jobs of first-term RMs probably differ depending primarily on whether the job is situated in a large or a small communications facility. Other differences might accrue dependent on the job site: that is, ship or shore installation. Therefore, hull types and shore installation types were both dichotomized by the panel into large and small categories.

Using the Enlisted Master Tape, the total number of RMs (both first-term and supervisor) assigned to each Navy facility categorized by the review panel was determined. Population totals in each of the four cells made by crossing facility size by sea/shore location was used to determine the number of questionnaires to be mailed.

Because the data to be collected that had the most variability consist of the proportions of personnel performing tasks, the simple random sampling formula for binomial data (with correction for finite populations) was applied to each cell. This ensured an adequate sample for each of the four potential job subtypes, should separate tests later prove necessary; that is, if a common core test can not be achieved.

In view of the fact that each cell includes a number of different types of ship or shore installations, the sample was drawn proportionally within each cell according to installation type. The PQ split used in the sampling formula was worst case (i.e., .5), the level of tolerance was .10, and the probability level set at .95. Thus:

$$N = \frac{N'}{1 + N'/\text{Population}}$$

The final questionnaire was sent to approximately 500 first-term RMs and 500 supervisors of RM personnel throughout the world. To facilitate timely return of the questionnaires, personnel assigned to ships that were deployed, on overseas cruises, or in overhaul status were eliminated. Excepting this restriction, sampling was random. The review panel asserted that there is no difference whatever in the RM job between overseas and CONUS locations. Therefore, we expect this exception to have no bias on sampling and consequently no effect on task selection.

The questionnaires were addressed to individuals by name, and delivered via the commanding officer. Completed questionnaires have been received and their information entered into a computertized data base. Figure 2 reflects total sent, returned, and usable questionnaires for each sampling matrix cell.

FIRST - TERM

	Large	Small
Shore	140 117 90	120 91 59
Ship	137 109 85	142 105 81

SUPERVISOR

	Large	Small
Shore	138 112 90	119 92 65
Ship	114 85 61	132 92 82

Figure 2. Questionnaires sent and returned[1]

Test Content Domain

While final analyses are not yet complete, preliminary data indicate that we obtained reliable judgments. PDRI researchers calculated intraclass correlation coefficients for the two major scales (frequency and criticality). These reliabilities were corrected using the Spearman-Brown formula. All were in the .90s. Similar calculations could not be done on the percentage performing or complicatedness scales because the absence of a rating for a task not performed or supervised could not be given a score. Thus, the frequency and criticality scales will be given emphasis in identifying and selecting the sample of tasks for item development, with less weight given to the scales pertaining to errors and complicatedness (which also showed more restricted variability across tasks).

To ensure that the job of the first-term RM is adequately represented in the final sample of tasks, functional job categories were superimposed on the set of 127 tasks. Final selection will be proportional to the number of tasks included in a category rather than from the entire list.

In order to derive the job categories a taxonomy of tasks was developed at a workshop comprised of 12 senior staff personnel drawn from the Navy's entry level RM training course in San Diego. Cards containing the 127 task statements were sorted into categories of tasks that are of the same type or that deal with the same type of equipment. The sort yielded the following categories: (1) Preparing and processing messages/Establishing communications; (2) Setting up equipment; (3) Maintaining equipment; and (4) Handling secure materials.

---

[1]Personal communication. Mr. S. Lammlein, PDRI, 23 October 1985.

The judgments made in the questionnaires will guide task selection. A successive hurdles technique will be employed to obtain the final job sample. First, the most critical tasks in each category will be selected. Then, tasks with low frequencies or low percent performing or supervising both overall and within each of the sampling cells will be replaced. Finally, tasks that have low values in percentage of performance errors or complicatedness will be replaced. Figure 3 summarizes the task selection procedure. Tasks will be subsumed within the four functional categories, with each category reflected in the final test composition by at least one task. Ultimately, the test content domain for the RM rating will be a sample of the 127 tasks, organized into functional categories.

1. Select tasks with overall most critical ratings

2. Replace any tasks with:
   a. low frequency of being performed
   b. low percentage of sampling populations performing
   c. low percentage of performance errors
   d. low complicatedness

3. Ensure proportional representation by category

Figure 3. Successive hurdles for final task selection

It is important to note that, at this point, none of the 127 tasks has been eliminated because it is unsuitable for hands-on testing. The complete list of tasks selected as a result of the survey will be presented to the review panel for its judgment as to suitability for testing in the hands-on mode. Unsuitable tasks, if any, will be replaced using the procedures used in developing the original sample. Subsequently, a hands-on test, computer-based simulation test, and set of rating scales will be developed.

Summary

Because no hands-on tests exist for measuring on-the-job performance, it was necessary to design and develop them. In turn, this required selecting critical tasks for which hands-on test items will be developed. Critical task selection was accomplished by an approach that combines supervisor and job incumbent questionnaire response through survey administration with quality control checks exercised through experience-based expert judgment. The result is a meaningful and manageable subset of tasks that can be converted into performance test items. The procedures used to obtain the job final sample ensure that the tests will have the highest possible content validity. We expect the results of critical task selection effort to materially contribute to the success of the Navy job performance measurement program, and ultimately to enhanced selection, classification, and assignment methodology.

Reference

Guion, R. M. (April 1979) Principles of work sample testing: III. Construction and evaluation of work sample tests (ARI TR-79-A10). Alexandria, VA: Army Research Institute for the Behavioral & Social Sciences.

Cognitive Predictors of M1 Tank Gunner Performance

Barbara A. Black
U.S. Army Research Institute-Fort Knox

Potential weapon system capability and achieved capability differ in large measure as a function of crew performance. To maximize Armor system effectiveness, the U.S. Army is committed to optimally selecting and training tank crewmembers. Within the four-man tank crew, specific emphasis has been placed on identifying soldiers who possess the requisite aptitudes and abilities to become proficient Armor tank commanders (TCs) and gunners. Early identification of these high-ability soldiers can lead to improvement in overall tank crew performance and in the cost effectiveness of training programs.

The realization of this goal is contingent on possessing the capability to measure abilities that have been identified as critical to job success and then being able to demonstrate the relationship between those abilities and actual job performance. The first step in this process is the analysis of job requirements for each position under consideration; the second step is the development of ability measures or tests. The U.S. Army Research Institute has for several years conducted research in job analysis and the development of tests for the prediction of Armor crewmember performance. This research has involved the development of tests for predicting success in basic and advanced individual Armor training, as well as in operational units.

Both job sample and paper-and-pencil tank gunnery predictor tests have been evaluated by ARI over a period of about ten years. The results of a meta-analysis completed on 15 of these data sets indicate that job sample tests were, across studies, better predictors of performance by job incumbents than were paper-and-pencil tests (Black & Campbell, 1982). Drawbacks to job sample testing do exist; they are similar to those identified from the psychomotor testing programs of the 1940s and 1950s: cost, increased administration time, and equipment unreliability (Melton, 1947). However, the advent of microprocessors and the increasing availability of high fidelity simulators may remove or reduce several of the major concerns in the use of job sample tests, specifically, the requirement for special equipment, the need for continuous calibration, and the difficulties involved in unit-collocated testing facilities. Job samples tests developed for incorporation into on-line or forthcoming unit-located simulators may improve the cost effectiveness of testing, reduce testing time requirements, and eliminate the need for special equipment apart from the simulator itself.

In addition, certain demographic variables have been found to correlate with tank gunnery performance across numerous studies for the past few years. These findings characterize the successful tank crew as being commanded by (a) a noncommissioned officer (NCO) with more time in the TC position than other TCs, (b) a TC who has trained longer with the gunner with whom he fired (Eaton & Neff, 1978), and (c) a TC

who has a history of having qualified crews (Piers & Sauer, 1982). None of these findings is particularly unexpected, but unfortunately, none is useful in the early identification of high-performing TCs. Yet this information is valuable in terms of providing data on variables whose covariance with the predictor measure may obscure the relationship of interest.

Initial efforts to evaluate predictors of crewmember performance in tank firing, driving, and loading used paper-and-pencil tests because they are the most cost effective and least time-consuming approach to performance prediction. Greenstein and Hughes (1977) used Armor trainees and limited their effort to the use of paper-and-pencil tests in the psychological literature or in use by the Army at that time. For example, they included Lauer's (1952) tests of Visual Memory and Attention-to-Detail, as well as the Armed Forces Qualification Test (AFQT) and three composites from the Army Classification Battery (ACB): Combat Operations (CO), Field Artillery (FA), and Motor Maintenance (MM). Significant correlations were obtained between the paper-and-pencil tests and loading errors and driving performance. None of the 11 paper-and-pencil tests in the study predicted tank firing scores. However, in a study commonly referred to as the Gideon report, Wallace (1982) presented the results of the 1982 European Canadian Cup trophy competition. He correlated the AFQT scores of tank commanders on the American team with their crew's live-fire gunnery scores and obtained a coefficient of .739 ($p < .01$, $N = 13$). The correlation between AFQT for the gunner and firing score was not significant. This research has prompted considerable interest in the existence and strength of relationships between the mental abilities of crewmembers and successful tank crew performance.

Scribner (1984) cited by Phillips (1985) in his report on Career Management Field (CMF) 19, found significant relationships between gunner/TC AFQT and tank Table VIII firing scores for 1131 tank crews training in Europe. This effort represents the largest database from which AFQT relationships have been obtained for live-fire performance.

While previous research indicates that certain testing techniques hold promise for Armor crewmember performance and more information is now available concerning important intervening variables, the availability of appropriate and useful criteria against which to validate predictor tests has remained a problem. Criterion measures used in past research include scores form live-fire gunnery exercises, Multiple Integrated Laser Engagement System (MILES) exercises, supervisory ratings, peer ratings, Skill Qualification Tests (SQTs), tests administered during the course of normal Armor training. Efforts to explain the inconsistencies found in past research have brought to light many disadvantages associated with the current job performance criteria available in Armor, especially those associated with gunnery.

Scores obtained from live-fire gunnery exercises often provide data that are not comparable between units or even between tanks. It is conceivable that with a company of tanks firing over a period of

several days, the condition of the weather, tank equipment, and range equipment could change to such a degree that no tanks fire the same engagements. In addition, for any specific tank, changes in ammunition characteristics, equipment performance, and firing conditions may reduce the reliability or increase the error variance for within-tank performance measures. Thus, low reliability of the criterion measure may have been a large contributing factor to the relatively inconsistent findings of past research.

In addition, it should be pointed out that tank gunnery tables are collective exercises. Engaging targets and measuring the results of those behaviors in such values as "time to engage" or "proportion of hits" produces a crew-level evaluation or, in the case of a Table IX exercise, a platoon-level evaluation. The relative contributions of individual crewmembers are difficult to ferret out. In fact, it is not uncommon for unit commanders who are short on high-quality personnel to pair mature, experienced TCs with novice or ineffective gunners to ensure that the tank crew will be rated "qualified." On the other hand, very effective gunners may find themselves in crews with ineffective TCs and fail to qualify their tanks during annual gunnery, thus making it virtually impossible to use the results of tank table exercises to make statements about individual performance.

The present effort involves the evaluation of a battery of M1 gunner performance prediction tests. Specifically it is an evaluation of the relationship between AFQT and the individual performance of gunners, using ratings from their supervisors, hands-on tests, and the crew/collective criterion, Table VIII gunnery. By virtue of the human factors engineering of the M1 tank system, the M1 gunner has more responsibility for fire control than the gunner on any other U.S. tank system. This increased responsibility, combined with the knowledge that from the soldiers selected and assigned as gunners come the future tank commanders (TCs) of the Armor force, makes it important to optimize the selection and assignment of M1 gunners.

METHOD

Subjects

The subjects, 123 M1 tank gunners representing four battalions, were selected for participation based on supervisors' ratings. Ratings were completed by company commanders and a senior NCO chosen by the commander. The two raters from each company were instructed to reach a consensus rank order for gunners, based on each gunners' demonstrated ability in performing gunnery-related tasks and on their availability for testing. Raters were asked to disregard gunners' performance in such nongunnery areas as military courtesy. In addition, raters were asked to consider the gunners' performance apart from that of their respective tank commanders or crews; that is, to rate gunners high if they were proficient, even though their crews may not have qualified on the most recent gunnery. Eight gunners were selected for testing from

each company. The four rated most proficient in each company and the four rated least proficient were tested. One company was exempted from the testing because of prior commitments. Thus two groups were formed, highly rated gunners and gunners receiving low ratings.

ASVAB. Four subtests from a research version of the Armed Services Vocational Aptitude Battery (ASVAB) were used to obtain estimates of soldier scores on the AFQT. Due to time constraints imposed on the overall testing process, a scaled-down version of each subtest was used. Subtests were shortened by randomly selecting 50% of the questions for administration. The AFQT consists of a combined score obtained from the following ASVAB subtests: Numerical Operations (NO), Paragraph Comprehension (PC), Arithmetic Reasoning (AR), and Word Knowledge (WK).

Tracking Test. This hands-on test used a snakeboard, an M1 tank, and an M55 laser boresighted with the main gun. A specially built device was used to pulse the laser automatically once per second for periods of 60 seconds. Soldiers were instructed to use the gunner station power control handles to track the snakeboard 12 times, 6 times from left to right and 6 times form right to left. To determine the gunner's accuracy on each trial the test administrator counted the number of laser pulses that fell on the snake. Speed was determined by recording the location of the final pulse, thus indicating the distance tracked during the 60-second trial.

M1 Computer Panel Test. Soldiers were tested on three operations of the M1 ballistic computer by means of a microcomputer-controlled simulation of the M1's computer panel. The simulator used a screen digitizer or touch panel placed over the face of a 12-in. color monitor. The operations consisted of enter - check data (ECD), and run computer self-test (CST). The software for the M1 computer test was developed to provide the soldiers with 3 instructional trials for each type of operation followed by 10 scored or test trials. The number correct on each operation and the time required to complete each test trial were recorded.

Criterion measures. The criterion measures included the supervisor's ratings and Table VIII scoresheets from the soldiers' most recent gunnery. Each gunner's overall Table VIII percentage was recorded as were the totals for day and night. Subtotals were also computed within both day and night scores for the moving main gun engagements.

Procedures

The validation effort was conducted double-blind; that is, neither the soldiers tested nor the test administrators were aware of the supervisors' ratings. The testing schedule allowed four soldiers to be tested in the morning and four in the afternoon. Sixteen days were required to test 123 gunners.

RESULTS AND DISCUSSION

Analysis of the data was completed in three steps: a) descriptive analyses by battalion, b) correlation between job sample measures and AFQT, and c) a break-out of performance by mental category.

Means and standard deviations for AFQT were computed for each battalion. No significant differences were found. However, significant differences were found among battalions for performance on the job sample tests: tracking and M1 computer panel. For this reason raw scores were converted to z-scores for all subsequent analyses.

Significant correlations between job sample measures and AFQT are presented in Table 1. It is interesting to note that AFQT correlated with number of hits for the hands-on tracking task (r = .360, p < .0001), indicating that gunners with higher AFQTs had more hits; that is, they were more accurate. Because tracking hits (accuracy) and tracking distance (speed) were negatively correlated (r = -.444, p < .0001), it is conceivable that gunners with a higher AFQT approached the tracking test with a greater emphasis on accuracy than did low AFQT gunners. Tracking distance did not correlate with AFQT. It is not possible to test the interaction between AFQT and the speed/accuracy tradeoff with these data.

Table 1
Job Sample Measures That Correlated with AFQT

Tests and Measures	r	(p <)
Tracking		
Number of Hits	.360	(.0001)
Composite Hits and Time	.316	(.0004)
M1 Computer Panel		
Enter-Check Data (ECD) Correct	.214	(.0172)
Enter-Check Data (ECD) Time	-.513	(.0001)
Enter-Check Data (ECD) Composite	.444	(.0001)
Computer Self-Test (CST) Time	-.339	(.0001)
Computer Self-Test (CST) Composite	.237	(.0082)

AFQT is more commonly reported in its grouped or categorized form, that is, by AFQT category rather than by percentile. When gunner performance on three job sample measures is presented by mental (AFQT) category the contrasts are striking. Note in Table 2 the drop in performance for category 3B and 4 soldiers. The measures presented include time and accuracy composites for the two M1 Computer Panel tests

and a General Ability Composite (GAC) formed by totaling the two Computer Panel composites and a time/accuracy composite from the Tracking Test. The GAC was developed using stepwise regression techniques; it was found to yield the most accurate prediction of AFQT. Average gunner scores were obtained by taking the average standardized score, adding a constant, then multiplying by 100. The correlation between GAC and AFQT was $r = .49$ ($p < .0001$).

Table 2
Average Gunner Score by AFQT Category

	AFQT Category				
Job Sample	1 (N = 13)	2 (N = 41)	3A (N = 27)	3B (N = 17)	4 (N = 25)
ECD Composite	198	141	139	47	19
CST Composite	170	151	91	55	10
General Ability Composite	346	211	105	26	5

The AFQT measures correlated with the high cognitive weighted tasks for gunner--ECD and CST on the M1 Computer Panel but not with the Table VIII scores. Because job sample tests were constructed with a combat criterion in mind, the snakeboard used in the hands-on tracking task involved tracing an extremely circuitous route, more similar to a threat target employing evasive maneuvers than to the slow-moving flank silhouette targets normally found on a Table VIII range. Thus one might conclude that if Table VIIIs were designed to more accurately mirror threat scenarios, the relationships between job sample tests or AFQT and live-fire gunnery might be easier to establish.

Conclusion

The present effort has involved the development and evaluation of skills tests for use as combat performance predictors. No relationships were observed for the supervisors' ratings and Table VIII day scores. These findings suggest that a UCOFT-implemented job sample testing approach may be more appropriate. UCOFT will allow high fidelity simulation of threat scenarios without the concomitant costs and safety hazards associated with a live-fire exercise.

Pending further research on job sample ability testing, a multiple hurdle approach to M1 gunner and tank commander selection may be suggested. Job sample ability testing for position-specific requirements may be combined with the on-site commander's evaluation of crewmember performance. This testing may offer a feasible approach to crewmember selection. Further research is required to assess how job sample testing can apply to tank commander selection.

A Retrospective Analysis of Instructional Technology Innovation Using General Systems Theory

Ana G. Ekstrom
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

The U.S. Military today operates in a worldwide, multi-lingual, multi-cultural context. The ramifications of this fact on linguistic training requirements presents a formidable challenge. There is a consistent need for a vast and constantly replenished supply of linguists trained in a wide range of languages with specialized skills to perform many and diverse tasks. Organizationally, the training requirement includes the capacity to teach even obscure infrequently needed languages as well as the ability to respond to the periodic, sudden and often unpredictable surges in demand for a particular language as military requirements develop.

The Defense Language Institute Foreign Language Center (DLIFLC), the foreign language training arm of the military, has found, despite claims made for various methods -- from Berlitz to Super Learning -- that learning a foreign language is a time-intensive process. Yet linguist personnel must be rapidly trained to functional levels of proficiency if the military is to benefit from short term enlistments, as well as meet surge requirements.

The use of interactive video in foreign language instruction has high face validity. Beyond the learning enhancement provided by multi-sensory presentation of instructional materials, interactive video offers the capacity to present a language in realistic situational and cultural contexts wherein the ability to understand and make computer responses, visibly results in simulated real-world consequences. Organizationally, a machine-based instructional system affords a versatile and flexible means of augmenting existing teaching structure, thereby potentially increasing institutional flexibility.

To take advantage of potential technological contributions to training capacity, DLIFLC established a specialized office to examine innovative instructional technologies. The New Systems Training Division (NST), identified the potential of the interactive videodisc in foreign language education and began preliminary exploration of developmental feasibility. This preliminary exploration led to what would ultimately become a complex "system" of individuals and organizations with the objective of field testing a Video Enhanced German Gateway Program (VEGG). *

* The term VELVET (Video Enhanced Learning Video Enhanced Testing) has been adopted by the developer for the video materials. The evaluation, concerned with assessing the use and impact of the materials on the German Gateway Program, rather than evaluating the materials themselves, uses the VEGG acronym.

This paper uses the evaluator's experience and knowledge of the VEGG system to illustrate the principles and utility of a systems perspective. This is not a report of findings and conclusions of a system evaluation. Rather it is intended as an example of how a systems perspective can assist in the organization and interpretation of events, actions, and reactions experienced in evaluation efforts even where the evaluation itself is not explicitly system oriented. We will begin with a description of the VEGG system in terms of its components and then illustrate the system principles described by Atwood (1985) using examples drawn from the VEGG system.

Initial development processes entailed feasibility and needs assessments related to materials development and the establishment of support within the military system to provide the requisite authorization, funding, and sponsorship base. The components of this particular instructional technology development system evolved incrementally with organizational units added in process. First requirements were for a "test bed", materials development personnel, equipment, and initial funding. The test bed requirement brought in the German Gateway Program (GG); materials development added two contractors; start-up equipment and materials development funding was provided by the Army Communicative Technology Office (ACTO) which became the technical sponsor; and test/evaluation funding was provided by the Training Development Institute (TDI) which became test sponsor. TDI subsequently tasked the Army Research Institute (ARI) with evaluation responsibility. TRADOC reorganization turned TDI into a new office of Training Technology Agency (TTA) and ACTO into Communicative Education and Training Systems Management Office (CETSMO) with consequent changes in personnel and uncertain sponsorship.

As the project moved toward the implementation phase, contractor and equipment funding were assumed by the National Security Agency (NSA) and only one of the two original contractors being retained. Within DLIFLC, quality assurance and research components were identified to review instructional materials and evaluation procedures respectively. Arrangements for video production were made with specialists from Ft. Dix and a full time Subject Matter Expert (SME) was provided to the project by the DLIFLC German school.

With the major VEGG system components now identified, we can examine the VEGG system applying General Systems Theory. The illustrations are necessarily drawn from the view available from the evaluator's window. Looking out the window of a different system component would certainly give a very different picture of the system itself as well as supply a whole different range of experiences from which to draw.

Synergism

Synergism is an outgrowth of regularized interaction, communication, and coordinated activity in the shared pursuit of system objectives. Time, distance, structural arrangements, expertise, and policies all impact on system synergy. Momentum in project development can be seriously affected by interruptions. The VEGG project experienced more than a one year delay between contractor selection and contract award for start-up activities and an additional one year postponement of field testing while awaiting completion of instructional materials. Such delays may divert evaluator and contractor focus and energies away from system objectives and into other projects, and reduce levels of interest, commitment, and involvement of sponsors and test units.

881

Conversely, delay may also serve positive functions by allowing participating units to accommodate to system requirements -- especially in the early stages of development. For example, the SME has a critical function in this system as the intermediary between developers and GG, translating language and instructional requirements into programmable lessons. The previously mentioned contracting delay provided the time needed for the SME to arrange to be released from teaching activities and to complete test development responsibilities. In this instance, the delay facilitated project development by enabling the SME to devote full time to instructional development when the project got underway. Likewise, the same period provided the evaluator (ARI) with the opportunity to become more familiar with the specialized areas of foreign language education and interactive video systems.

Ideally from a synergy point of view, core development staff (and perhaps evaluator staff) would share a common central facility. Where this is not possible, means of communication and feedback which simulate this type of working environment need to be established. The VEGG project began development efforts relying on the mails, phone calls, and periodic site visits back and forth to coordinate DLIFLC and contractor efforts. The contractor sent completed lessons to DLIFLC for review. Feedback, corrections, and suggested changes were phoned or mailed back. Prompted by something of a crisis created by the delay in materials development, direct telephonic computer links were established between the contractor and NST enabling instant feedback, correction, and literally simultaneous developmental effort to occur.

System Structures

Lines of authority and division of responsibility define vertical and horizontal structure respectively. When control of decision-making and resource allocation resides within the development system, the system is able to organize to achieve its objectives. When major control does not reside in the development system, it can only be understood as a sub-system within a larger system(s). Under the latter condition, the development system's objectives may become secondary to primary objectives of the larger system(s). Resources then must be negotiated and participants may be subject to conflicting authority.

ARI's original relationship to the VEGG project was on a Technical Assistance Service (TAS) basis which is by definition a short-term, limited resource effort. This TAS-level relationship specifically defined ARI as a limited stakeholder in the VEGG system. The long-term intensive commitment required for a systematic systems evaluation is precluded - - the VEGG system's evaluation resources could be diverted at any time into ARI's primary program objectives. In addition, ARI evaluation staff is subject to both the authority of the ARI structure in determining evaluation design features, procedures, and time commitments and to the German Gateway management in defining permissible data collection intrusions into the program.

By virture of his position as intermediary for a number of components of this system, the SME is even more enmeshed in multiple lines of authority. As translator of foreign language pedagogy into technology based instruction, the SME may be caught between the conflicting policies or needs of the various

components of the system. For example, GG instructional policies demand exclusive use of German in educational materials. This GG policy came into direct conflict with an NST/contractor decision to use limited English to support students' initial use of the IVD, resulting in directly opposed demands on the SME. This illustration exemplifies a type of technical issue that emerges in the course of converting conventional instructional methods to media-based technologies.

As evident from the above, lines of authority determine where critical decisions are made, by whom, and in terms of what objectives. Depending on their impact on system relations, decisions also promote, build on, inhibit, or destroy system synergy.

Stability

The ability to achieve a degree of internal stability is required to function in terms of system goals. At the same time, instructional technology development systems must overcome stability in the existing educational structures in order for technological innovations to be adopted. The VEGG project faces both challenges.

The process by which the VEGG system was constituted was essentially additive. That is, units were added as required by either the system's purpose or as imposed by external authority. The result of this process is something akin to a perpetual motion system forever adjusting to the requirements of new components. To illustrate from the evaluator's perspective, ARI, in its evaluation effort, identified the inadequacy for evaluation purposes of the existing student performance measures used in the GG program. This necessitated the development and administration of a new test, a requirement which, when combined with data collection needs, significantly added to the responsibilities of teaching and testing staff of the GG program. When the contractor added evaluation staff approximately half way into the baseline data collection (contractor funding actually occurred after data collection had begun), weaknesses in evaluation measures were identified by their foreign language specialist evaluator which imposed additional requirements on ARI and DLIFLC system components. Still later, a DLIFLC researcher entered the system (newly hired to fill a newly created position) and identified additional contractor evaluation responsibilities. The disruptive influence of these events might have been averted if an evaluation component, operating from a systems framework in the early stages of development, had facilitated a broader more realistic perception of the nature and requirements of the overall program.

A systems oriented evaluation also would have broadened the scope of outcome concerns to include examination of processes and relations which facilitate adoption of technology once developed. On the VEGG project two contractors were involved in preliminary feasibility and needs assessment activities. One contractor, with both technical and linguist capability, developed a product - - an interactive videodisc demonstration lesson. The other contractor, lacking linguist capability, operated as a consultant, training members of the DLIFLC staff in design concepts and assisting them in developing sample lesson designs. From the larger system perspective, the product is the desired outcome. Products can be more efficiently acquired from a contractor than by expending limited and essential specialized German

instructional staff resources in product development.

The selection of the first contractor established a "product model" (Rutt, 1984) relationship between the contractor and NST/GG in the development system. This selection limited the demands made of GG staff to that of insuring product quality/acceptablity for their program. This decision reduced both the potential strain on DLIFLC German staff resources and the dependence of the developmental system on those resources. At the same time, it reduced the GG stake in the system. The second contractor, requiring DLIFLC German language and instructional expertise, would have necessarily involved the German staff to a far greater extent. This involvement might have fostered greater appreciation for developmental issues and strengthened investment in the final product and its use in the program.

Optimization

"If sub-systems operate to optimize their own individual performance the net result will almost never be overall system optimization" (Atwood, 1985). Several examples from the VEGG system illustrate this point.

In the case of the previously described successive efforts to modify the evalution design, one component proposed modifications in the evaluation design which would have strengthened the evaluation but would have expanded data collection efforts. The outcome was that the system collectively operated to limit expanded data collection due to the excessive intrusion and demands that would have been placed on GG student and staff resources, ARI evaluation resources, and SME time.

As developers discover the advantages offered by various technological possibilities, additional time is required to integrate new features into product development. For example, VEGG developers became aware of the highly desirable possibility of incorporating audio commentary or questions into their materials through use of the second audio track of the video disc. Understanding the full potential of a technology options is an evolutionary process. That evolutionary process, combined with the constant marketing of new hardware, software, and courseware that occurs in the electronic media field may entice developers into exploring new or newly recognized possibilities at the expense of timely product completion.

Another example deals with quality assurance review of instructional materials. This must necessarily be a constant process, deciding on almost a frame-by-frame basis what is excellent, what can be lived with, and what is totally unacceptable. Grammar and syntax errors are easily recognizable as unacceptable. Reviewers tend to readily identify excellence. When video instructional materials fall into the no-man's-land between the extremes, constant and difficult decisions must be made in terms of what can be lived with. Insistence on excellence only, may result in no product at all.

The fine art of negotiating between "must have/do" and "would like to have/do" characterizes the process of achieving system optimization over optimization of individual component performance. Systematic documentation of this negotiating process would provide information relating to the balance that exists among the array of sub-system objectives and between sub-system and development system objectives.

884

Equifinality

The principle of equifinality has two important implications for system evaluation. The first relates to the need to identify options and alternatives for "as the number of alternatives increases, the system is more able to adjust to unforseen changes in the environment" (Atwood, 1985). As has been implied throughout this paper, there are many ways to achieve system objectives. Different choices in terms of equipment, contractor, evaluation approach and design, target programs and language would have substantially changed the nature and operation of the VEGG system and its final products. Decisions made in the early stages of development limit what can follow and determine to a great extent what must follow. For example, the choice of a Gateway program limited the test population to officers, dictated a baseline/field test comparison rather than an experimental design, and limited outcome performance measures to short term learning at low level, limited range proficiencies. These limitations may well be less significant to overall system goals than the advantages gained for example from the greater liklihood of product adoption by an already equipment-based (audio cassette) program and the limited but more manageable scope of evaluation. The important point is that the involvement of all parties early in the planning process and in decision-making throughout development increases the system's ability to anticipate its overall needs and to develop contingency alternatives.

The principle of equifinality also maintains that the objectives of a system may be achieved "from a number of different starting conditions and through a variety of means" (Atwood, 1985). Fortunately, a system's ability to achieve its purposes does not depend on a precisely defined set of components, processes, and relationships. However, within the framework of equifinality, the cumulation of a body of information - - through systems evaluations of a variety of instructional development systems - - would begin to refine our understanding of what works better, when, and how - - and not at all.

To the extent that a systems approach to evalution does not result either in information overload or in surfacing such an overwhelming number of issues that instructional development initiatives are stymied, formative systems evaluation enhances the development process. Properly used and understood, the systems approach enables evaluators to perform a comprehensive job, collecting information that can be used both internally and externally to achieve system objectives.

References

Atwood, N.K. (1985). A Systems Approach to Evaluating High Technology Training. Paper presented at Military Testing Association (27th Annual Meeting).

Churchman, C.W. (1968). The Systems Approach. New York: Delacorate Press.

Rutt, D.P. (1984). Consultation in Instructional Development: A First Look. In R. Bass and C. Dills (eds) Instructional Development: The State of the Art II. Dubuque, Iowa: Kendall/Hunt. 294-309.

von Bertalanffy, L. (1968). General Systems Theory. New York: George Braziller.

Wolf, W.C. (1984). Linking Knowledge Production and Needs of Knowledge Users. In R. Bass and C. Dills (eds.) Instructional Development: The State of the Art II. Dubuque, Iowa: Kendall/Hunt. 259-379.

A Systems Approach to Evaluating High Technology Training

Nancy K. Atwood

U.S. Army Research Institute Field Unit
Presidio of Monterey, California

Military training demands have steadily increased with the advent of the computer age and the development of complex electronics and weapons systems. At the same time, the volunteer force has been faced with a constrained supply of skilled recruits. To address these dual problems, instructional applications of new technologies have received considerable attention in the military training community as a strategy for leveraging training resources (O'Neil, 1981). For example, the U. S. Army Research Institute maintains a "technology watch" to identify potential techniques for increasing the cost effectiveness of training, for raising levels of learning at the same cost as conventional methods, or for achieving the same results as conventional methods at lower cost.

One promising new technology is "Intelligent" Computer Assisted Instruction or ICAI. ICAI, a field within the larger discipline of Artificial Intelligence, draws upon emerging computer technologies and cognitive psychology in an effort to provide a new sort of computer-based learning environment. The intent is to provide an intelligent tutoring system that simulates a human coach or tutor. The enthusiasm of ICAI developers is reflected in the comments of Anderson and Reiser (1985) who argue: ". . . The prospect is great of providing every student with the educational benefits of a private human tutor. When this happens, the consequences for American education will be nothing short of revolutionary."

While ICAI is a rapidly evolving field, functional programs are appearing, funded in large part through the investment of DoD resources. It is vitally important at this time for the military training community to critically examine instructional applications of the new, AI-based computer technologies. This examination should be directed toward understanding the state of the art of ICAI technology and the process for developing ICAI systems with a view toward shaping the development process and the emerging product to military training needs and requirements.

This paper presents a strategy for designing and conducting early or "formative" evaluation of instructional applications of new technologies. The approach, based on General Systems Theory, is first described. This conceptualization was in fact stimulated by problems experienced while attempting to apply a conventional evaluation methodology to three ICAI projects (Baker & Atwood, 1985). Then insights into the utility of the systems approach are presented based on a post-hoc analysis of experience with these projects.

A Systems-based Formative Evaluation Strategy

Virtually all prescriptive models of the instructional development process specify the need for "formative evaluation" to identify program strengths and weaknesses for program improvement by collecting information on instructional processes and learner performance. A case in point is the Instructional System Development (ISD) model of military training development which

prescribes evaluation as an integral component of the development cycle (O'Neil, 1979). However, evaluation is generally conceptualized as one step in a linear sequence of development procedures which is initiated when a prototype product is ready for field-testing.

In contrast, the systems-based formative evaluation strategy to be described here conceives of evaluation as an ongoing process that begins with project start-up and continues throughout full project implementation. The approach views the organizations involved in instructional technology development as sub-systems within a larger system. The principles of General Systems Theory are applied to derive features of an evaluation strategy that will facilitate successful development and implementation of the technological innovation.

Systems theory encompasses a vast range of issues and approaches ranging from systems philosophy to development of mathematic system theories, to empirical research on system behavior, to systems engineering (Mattesich, 1982). However, at the heart of systems thinking is an emphasis on input-output features, a purposeful orientation with concerns for means-ends relationships, and feedback loops within the system to adjust the behavior of system components to yield the desired end-behavior of the entire system. (See, for example, von Bertalanffy, 1968.) The essence of systems thinking lies in a wholistic, non-linear view that constantly relates components of the system back and forth to the overall system and tries to reconcile their often conflicting goals.

Figure 1 presents a prescriptive systems model of instructional technology development. Generally, there are at least three interacting organizations or system components in the military training setting. These include the contracting agency that funds the project, the research and development organization (usually a private business or university) that develops the product, and the school that is the targeted user. (As shown in the figure, there are also sub-systems operating within each of these components.)

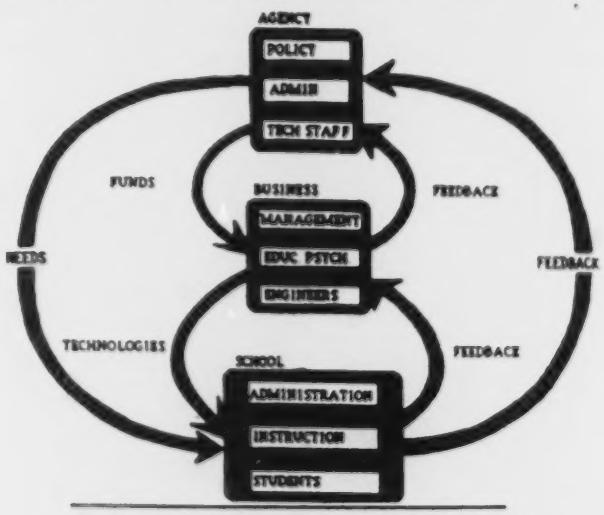


Figure 1 Systems model for instructional technology development

Traditionally, the development process has relied on minimal interaction between the above organizations. Given the natural bureaucratic barriers to

communication, organizations generally take responsibility for their portion of the process, with little communication and coordination with the others. Furthermore, organizations often have competing or conflicting goals (e.g., funding agencies looking for specific training applications, R & D organizations concerned with producing scientific knowledge, and user organizations looking for affordable and feasible enhancements to current procedures). As the history of Computer-Assisted Instruction (CAI) has amply demonstrated (Montague & Wulfeck, 1984), the end product often is not well targeted to the school's instructional needs, may be incompatible with the organization of the instructional program, and further may not be fully implemented or even used because those school officials having authority do not feel ownership or commitment to the project.

In contrast, the system view in Figure 1 recognizes the critical importance to technology development of communication through feedback loops between the various groups within the system. These loops provide a vital mechanism for ensuring that the resulting innovation is both objectively valuable and psychologically valued by the organizations involved with the project.

Principles drawn from General Systems Theory provide insights into conditions that will maximize the liklihood of achieving the overall goal of the system (in this case, successful implementation of an effective instructional technology innovation):

1. Synergism: As integration among system components increases, synergistic effects (i.e., synchronized energy) are more likely to occur. Extensive communication among system components, in this case, the funding agency, R & D organization, and school, are needed to develop the strong, well defined interrelationships required for integration to occur. This dialogue must yield mutual understandings and acceptance of relative roles and working relationships.

2. System structure: The clearer the vertical and horizontal structure, the more the system is in control and the more persons and activities within the system are vested with purpose and meaning. Again, the communication and feedback loops within the system properly used can serve to clarify the system structure and yield common understandings of the respective responsibilities of components within the system.

3. Optimization: If sub-systems operate to optimize their own individual performance, the net result will almost never be overall system optimization. Thus, in order to achieve overall success, system components must embrace the overall system goal, agree to cooperate, and devise strategies for accomplishing unique organizational goals within the larger context of the system goal. Joint planning sessions with mutual discussion and negotiation on project parameters can serve to create a sense of ownership by all parties and to adjust project plans so that the interests and needs of component organizations are incorporated.

4. Stability: All systems tend to have mechanisms for manning a steady state. This tendency leads to the well-recognized resistance to instructional innovations (e.g., Berman & McLaughlin, 1977) and implies a need for planning to anticipate the problem. Such planning should establish organizational supports for re-establishing stability after the introduction of the innovation.

5. Equifinality: The final state of an open system can be reached by starting from a number of different initial starting conditions and through a variety of means. As the number of alternatives increases, the system is more able to adjust to unforeseen changes in the environment and to accomplish its purposes. Thus, project planning should be open to multiple options for development and implementation to increase the adaptive flexibility of the effort.

The above systems view implies that formative evaluation should be structured to facilitate system functioning in order to achieve the system goal, to contribute to communication and feedback among system components, to provide information on the needs and concerns of groups within the system, and to gather information on program functioning and outcomes for revision purposes. Ideally, the evaluation should:

- be conducted by a team external to the system. An independent team maximizes objectivity and minimizes self-involvement in the interests of organizations within the system. As such, it can serve to facilitate overall system functioning.
- examine the functioning of system components as well as their interaction process from project development through full implementation. Recommendations for improvement should focus on promotion of cooperation among system components to achieve overall project goals.
- capture ongoing communication and feedback mechanisms between all system components. Description and identification of strategies for improvement are critical since these mechanisms can serve to increase integration and synergism of system components and to clarify system structure.
- address specific information needs of all organizations within the system. This information can serve to meet the needs and goals of individual components within the system and to resolve conflicts that may interfere with achieving the overall system goal. Explicit recognition of the needs of each organization also serves to establish the credibility and utility of the evaluation to all concerned parties.
- be sensitive to the stage of development of the project. Information needs and concerns change as a project matures and the freedom to make changes diminishes. In the early stages of project planning, focal points should include articulation of instructional/training theory, operational plans for product development, initial implementation planning, and use of communication/feedback loops within the system. In the middle stages of project design and development, focal points should include product design and development, theoretical fidelity and quality of the instructional process, preliminary implementation (with supports reestablishing system stability), outcome measurement and ongoing communication/feedback processes. In the later stages of project implementation, focal points should include outcome assessment (both cognitive and affective), measurement of side-effects, project implementation, cost-effectiveness, and ongoing use of communication/feedback loops.

- capitalize on multiple measurement methods for collecting information. Diverse methods, both quantitative and qualitative, are helpful in acquiring the full range of relevant information.
- yield a full documentary base to inform subsequent decisions and to provide a full record of intentions and accomplishments. This record will contribute to an understanding of the functioning of the system and its components and will provide a vehicle for discussion in subsequent project planning and evaluation activities, including follow-on projects.

Insights into the Need for a Systems Approach

The systems-based evaluation approach described above was used to analyze experiences and problems encountered conducting evaluations of three ICAI projects. The projects differed in their stage of development from early to middle to late. The project in the early stages of development was designed to teach maintenance technicians to troubleshoot and repair a complex electronic machine. The project in the middle stages of development was intended to assist novices learn to program in PASCAL. The project late in its development cycle was directed toward teaching basic mathmatics and strategic thinking skills within a game context.

The analysis suggested that the systems-based evaluation approach was a useful tool for understanding the problems that emerged with conventional formative evaluation methods and for identifying promising strategies for future evaluations of such technical innovations. Foremost, strong incentives are required for full participation by all groups within the system. While the funding agency and the school generally perceived participation in evaluation activities as in their own self-interest, project developers were considerably more skeptical and resistant to participation; evaluation activities tended to be perceived as directed toward uncovering weaknesses and diminishing their control over the development process. Certainly one strategy for gaining full cooperation from the developer is to include external evaluation in the project funding agreement.

Second, the relevant range of expertise needs to be represented on the evaluation team in a manner that is credible to the funding agency, the developer, and the school. In one evaluation study, only limited cooperation was obtained from developers due, at least in part, to the fact that the evaluators belonged to a different professional community than the developers and were not viewed as knowledgable or capable of providing useful information. For ICAI projects developed for military training applications, the relevant range of expertise minimally includes: instructional psychology, military training, artificial intelligence, and the subject matter to be taught (e.g., electronic maintenance, PASCAL programming). In addition to constituting an evaluation team with skills in these areas, additional credibility would be accrued by convening a technical advisory panel of esteemed colleagues of developers (i.e., key figures in Artificial Intelligence).

It is also vital for conditions to be established that foster the development of informal networks and the growth of trust and credibility from the inception of the project. In these evaluations, trust and credibility developed slowly over the course of many months and was sometimes impeded by lack of knowledge about strategies and events occurring earlier in the development cycle. Supporting conditions for fostering productive working

relationships include: involving all parties (including the evaluation team) at the beginning of the planning process and recognizing the need for long-term (i.e., multi-year) development and evaluation agreements. Such agreements are necessary for adequate communication and feedback to occur, for the extended development cycle encountered in projects with a high degree of technological complexity to run its course, and for mutual trust among individuals and groups to develop.

Finally, safeguards on the use and distribution of information collected by the evaluation team are required. Much of the difficulty encountered working with developers, and to a lesser extent, project implementers, arose from concerns about: (a) how and to whom the evaluation information would be disseminated; (b) whether and at what point they would have an opportunity to react to evaluation findings and to voice their interpretations of findings; and (c) how the evaluation results would be used. Thus, it is critical that all parties are assured that information will be treated appropriately and sensitively and procedures are in place to protect such assurances.

One such procedure for restricting information flow to guard such assurances is for the evaluation team to limit dissemination of its findings to organizations within the system, i.e. the funding agency, the R & D organization and the school. Groups should be provided an opportunity to react to the findings and to present their interpretations to the evaluation team. While ultimate incorporation of such input into the final evaluation report should be left to the professional judgment of the evaluation team, organizations should be provided the opportunity to prepare a "minority report", if they have serious disagreements with the final evaluation report. Finally, all publication and presentation of results to the technical and training community should be approved by organizations within the system studied.

It is hoped that a systems-based formative evaluation strategy will support the development of high quality instructional technology innovations, and, perhaps more ambitiously, yield a record that may provide guidance for establishing general policies related to training technology R & D.

References

Anderson, J. R. & Reiser, B. J. (1985). The LISP tutor. Byte, 159-175.

Baker, E. L., & Atwood, N. K. (1985). Shaping the wind: Formative evaluation of intelligent computer-assisted instruction (ICAI). Paper presented at Annual Meeting of American Educational Research Association, Chicago, IL.

Berman, P. & McLaughlin, J. W. (1977). Federal programs supporting educational change, Vol. III: Implementing and sustaining innovations. R-15891 8-HEW. Santa Monica, CA: Rand.

Mattessich, R. (1982). The systems approach: Its variety of aspects. Journal of the American Society for Information Science, 383-394.

Montague, W. E. & Wulfeck, W. H. (1984). Computer-based instruction: Will it improve instructional quality? Training Technology Journal.

O'Neil, H.F., Jr. (1981). Computer-based instruction: A state-of-the-art assessment. New York: Academic Press.

O'Neil, H. F., Jr. (1979). Issues in instructional system development. New York: Academic Press.

von Bertalanffy, L. (1968). General system theory. New York: George Braziller.

An Analysis of Attitudes Toward Instruction Among Vocational Education Instructors

George M. Usova, Head
Analysis and Evaluation Division
Shipyard Instructional Design Center, Atlantic
Portsmouth, VA 23709

Identifying attitudes toward instruction is important in determining program changes. An extensive study, conducted by George M. Usova sought to determine the instructional beliefs and practices held by skilled trades training personnel in naval shipyard apprenticeship programs.

The study surveyed the attitudes and practices of 256 key training personnel in apprentice trade training areas. The trades participating in the study were air conditioning and refrigeration, boilermaker, electrician, electronics, fabric worker, machinist, insulator, painter, pipefitter, rigger, sheetmetal, shipfitter, shipwright, and welder.

The trade trainers were asked to rate their feelings of agreement or disagreement toward educational practices by using a five point scale. The 17 item questionnaire measured attitudes on apprentice ability in physical and basic communication skills, use of self-paced and computer-managed instruction, adapting instruction to learner differences, and values held about inservice training. In addition, the questionnaire asked the trainers to indicate their primary method of instructional delivery, state how often they received inservice training, and tell how often apprentices used effective learning strategies to learn subject area content.

Background. Since the U.S. Navy is the single largest employer of apprentices in the nation, it has a substantial stake in assuring that its apprentice training programs are effective. One area of concern and investigation lies in the area of instructor training. Apprentices are trained by instructors who receive only 40-80 hours of preparatory instructor training.

The intent of the study, therefore, was to assess the attitudes and practices toward teaching and instruction among vocational trade instructors and other key training personnel in apprenticeship programs in naval shipyards. The importance of the study was to provide data from which to make better decisions for improving apprenticeship programs and identifying instructor training needs for inservice training.

Method. The Training Information Survey was developed as a two-part questionnaire to assess particular training attitudes and practices of training personnel (Group Superintendents, Shop Superintendents, Supervisory Training Instructors, and Trade Theory Instructors) involved in apprentice training at the naval shipyards (see Appendix A). The survey questionnaire is composed of two parts: (1) Part I: an attitude section that assesses agreement or disagreement with statements about apprentices' abilities, educational practices, and inservice training and (2) Part II: an "actual practice" section that assesses the extent and frequency with which instructional practices actually occur. Additionally, the questionnaire solicited biographical information (i.e., the respondent's trade, shipyard, and job position) that was matched with the questionnaire data.

In total, 472 questionnaires were sent to eight shipyards with 286 questionnaires returned (within the five-month period), which yields a return rate of 60.6 percent. A breakdown of useable questionnaires by respondent groups reveals the following figures: 19 Group Superintendents, 77 Shop Superintendents, 32 Supervisory Training Instructors, and 128 Trade Theory Instructors. The trade areas surveyed were air conditioning and refrigeration, boilermaker, electrician, electronics mechanic, fabric worker, machinist, insulator, painter, pipefitter, rigger, sheetmetal mechanic, shipfitter, shipwright, and welder.

Findings and Recommendations. As a result of the study, 11 separate findings were made from a data analysis of the study as a whole. The findings and recommendations incorporate the results from Parts I and II of the questionnaire.

1. Apprentice hires do appear to possess the manual dexterity and coordination skills necessary to meet the demands of skilled trades training.

2. Apprentice hires do lack to a substantial degree basic skills in reading, mathematics, and communications to meet the demands of trade theory courses. Nearly 70 percent of training personnel believe apprentices are deficient in one or more basic skills. Verbal comments additionally support this finding.

3. Self-paced instruction is not well-accepted as an instructional delivery mode. This finding is further reinforced through other data showing that 82 percent of trade theory instruction is group-paced. Group-paced (predominantly lecture) is a traditional form of instruction that nearly everyone is familiar with. While group-paced instruction can be effective in achieving learning, it is widely recognized and established that there are a variety of alternative instructional methods (self-paced, individualized, simulation, grouping, team learning, and others) that can be more effective in achieving certain learning outcomes. It is the awareness, knowledge, and implementation of varied methodologies that can strengthen the instructor's teaching competency.

Recommendations. It is recommended that, through inservice training, training personnel be shown the educational values and benefits that can arise in student learning through self-paced instruction. The point here is not to attempt to force self-paced instruction upon trainers nor is it to be implied that self-paced instruction is the best delivery mode; rather, the intent is to recognize the use and role that self-paced instruction can play as a part of an eclectic (varied) approach to training.

4. There is wide agreement and support that a planned on-the-job instructional program could benefit shop training. Support, therefore, exists for a program that systematically permits and ensures that critical job skills are experienced and developed during on-the-job training.

Recommendations. It is recommended that all trade apprentice programs ensure that apprentices experience critical job skills through a planned sequence of experiences.

5. Overall, there is uncertainty over whether computer-managed instruction would assist learners.

Recommendations. As stated in recommendation number three, there is a need to educate trainers to the benefits inherent in using a computer-managed system. In an age of information explosion, instructional technology, and cost/time efficiency requirements, computer-managed instruction can perform an integral role.

6. There is widespread acceptance to the need for hands-on training removed from the production environment. This educationally sound concept permits learners to practice and develop skill proficiency on training aids, mock-ups, and simulators before actually "working a job". In this manner, theory and practice are closely tied together to ensure skill mastery and thereby reduce the possibility of costly errors that might occur on-the-job.

Recommendations. It is recommended that all trades explore additional sources and acquire, as needed, training aids, mock-ups, and/or simulators to develop improved hands-on training for vestibule training.

7. There is overwhelming support for the value, benefit, and need for instructor inservice among all trade training personnel. This belief lies in the generally accepted concept that if instructors become more competent in the area of training (learning theory and practice, objective and test-item development, instructional strategies, etc.), then students will achieve in greater proportions. Central to the issue of inservice training is the type and quality of that training. Inservice training must be directly responsive and relevant to the improvement of the teaching-learning process, i.e., the topics of inservice must address instruction-competency areas.

Recommendations. It is recommended that shipyards develop a systematic instruction training inservice program that is based upon instructor needs.

8. There is strong acceptance for the use of modern audiovisual aids to support instruction.

Recommendations. It is recommended that trade training personnel review their audiovisual acquisitions to determine whether they possess the proper type and amount of equipment necessary to support and facilitate learning of trade theory courses.

9. At the present time, strong agreement exists that shop trade theory courses meet the learning needs of apprentices.

10. Uncertainty exists about whether courses are presented in the different ways in which students learn. Supporting this uncertainty belief is survey data which show that 70 percent of respondents believe that instructors only "sometimes" adapt instruction to learner differences. Again, this is indicative of instructional rigidity or inflexibility to alternate methods and parallels the finding stated earlier: 82 percent of shop instruction is group-paced.

Recommendations. Launch a program of instructor inservice which includes topics on detecting and diagnosing learner abilities, individualizing instruction, classroom grouping patterns, peer learning techniques, and other topics as cited in recommendation number three.

11. Only 22 percent of students in trade theory courses demonstrate efficient learning strategies and study skills. Learning strategies and study skills, such as notetaking, listening, time management, memory techniques, and study methods, are the means through which to acquire and retain information. Responsibility for teaching those learning skills is a dual responsibility, requiring effort from both student and instructor. The instructor is responsible for showing students how to learn and better recall that specific information that is presented in the classroom; again, the instructor can be taught how to do this through inservice training. The student, on the other hand, can be taught the general application of learning strategies and study skills and must be personally responsible for applying those skills.

Recommendation. It is recommended that a course on learning-study skills be provided instructors as a component of an inservice training program and that a separate and distinct program on learning-study skills be offered to all apprentices during their first year in the program. Such a two-pronged approach will maximize learning efficiency.

TRAINING INFORMATION SURVEY

Trade _____ Shop _____

Status: (Please check the one that applies to you.)

_____ Group Superintendent
_____ Supervisory Training Instructor
_____ Shop Superintendent
_____ Trade Theory Instructor

The Shipyard Training Support Center role is to support your effort to strengthen and improve apprenticeship training. We need your assistance in completing this survey form to help us define and clarify shipyard apprentice training needs for your shop. Please take a few mirutes of your time to respond to the following statements. Return the completed form to your employee development representative, or if requested, your group superintendent. Thank you.

Directions, Part I

After reading and considering each statement, assign the number that reflects your feeling about the statement.

1 - Strongly Disagree
2 - Disagree
3 - Undecided
4 - Agree
5 - Strongly Agree

1. ___ New apprentice hires have the manual dexterity and coordination to meet the demands of skilled trades training.

2. ___ New apprentice hires possess the necessary basic skills in reading, mathematics, and communications to meet the demand of skilled trades training and trade theory courses.

3. ___ Self-paced instruction (that allows each apprentice to control his own rate of learning) could benefit the learners of skilled trades training in the shop.

4. ___ A planned on-the-job instructional program could benefit training in the shop.

5. ___ Computer enhanced and computer managed instruction would enhance the learning potential of students in the shop.

6. ___ Hands-on training, removed from the production environment, is needed to increase trade skills.

7. ___ Shop instructors could profit from an ongoing program of in-service training designed to improve teaching skills.

8. ___ The improvement of instructor teaching skills will improve student learning.

9. ___ Current and modern audiovisual aids and equipment are important in supporting student learning in skilled trades training.

10. ___ Presently, shop trade theory courses meet the knowledge and skill learning needs of apprenticeship students.

11. ___ Presently, shop trade theory courses are presented to meet the different ways in which people learn.

Directions, Part II

In order for us to more effectively understand your shop's training program, please respond to each statement as it relates to your shop's operation by checking ( ) the appropriate blank. Thank you.

1. Most of the instruction delivered in my shop is

   _____ group-paced _____ self-paced
   _____ others; specify _____

2. Instructors regularly receive in-service training in the improvement of teaching skills

   _____ more than once per year _____ once per year
   _____ none received
   _____ other; specify _____

3. Records kept on student progress/skill levels are

   _____ maintained regularly _____ not maintained
   _____ maintained sketchily

4. As a group, instructors adapt instruction to account for differences in learner abilities, ex., slow learners vs. fast learners

   _____ always _____ sometimes _____ rarely

5. As a group, students in trade theory courses demonstrate skill in efficient learning strategies and effective study skills

   _____ always _____ sometimes _____ rarely

6. As a group, apprentice hires are lacking the necessary basic skills in (check as many that apply)

   _____ reading _____ mathematics _____ communications
   _____ other; specify _____

Direction, Part III

Please use the space below to comment further or explain any of your responses. Thank you.

Please return the completed survey to your employee development representative, or if requested, your group superintendent.

RELATIONSHIP OF AN EXPERIMENTAL HISPANIC ENLISTMENT SCREENING TEST TO AFQT

John J. Mathews, Naval Training Systems Center

Carla M. French, University of Central Florida

Efforts are underway to increase the representation of Hispanics in the military. These efforts are not only founded on affirmative action goals but also in response to a decreasing pool of eligible youth. A large number of prospective recruits are excluded from military service due to low scores on the Armed Services Vocational Aptitude Battery (ASVAB). The potential aptitude level of many applicants is underestimated due to limited English skills of some recent immigrants and members of Hispanic subcultures. An aptitude test (Hispanic Enlistment Screening Test) written in Spanish has been developed by the Navy Personnel Research and Development Center (NPRDC) as an aid in identifying Hispanic youths whose ASVAB scores would likely be acceptable after remedial English training. This study is a preliminary evaluation of this test in the context of the Navy's Accession English Language Training (AELT) program. Participants are currently limited to high school graduates with English Comprehension Level Test scores at the marginal level of 45-80. They receive up to 22 weeks of AELT at the Defense Language Institute (DLI) prior to ASVAB retesting.

METHOD

The subjects were Puerto Rican males who completed AELT and were retested on ASVAB by April 1985 (N = 205). Another 112 recruits were excluded from this study because they were still in training or had been discharged prior to ASVAB retesting. Predictor variables included scores on the experimental Hispanic Enlistment Screening Test (HEST) and the Armed Forces Qualification Test (AFQT) composed of subtests from the ASVAB. Like AFQT, the HEST has four subtests--Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), and Numerical Operations (NO). The ASVAB was administered to the subjects prior to Navy enlistment in Puerto Rico. The HEST was given at the DLI upon AELT entry. Those completing AELT (2 to 5 months) returned to Orlando to begin recruit training. Early in training, the subjects were retested on the ASVAB.

It is expected that as a result of AELT, posttraining AFQT scores, particularly on verbal subtests, will be higher than pretraining AFQT scores. The main hypotheses to be tested are that HEST scores will correlate higher with post-AFQT scores than will pre-AFQT scores and that HEST scores will add appreciably in predicting post-AFQT scores. Hispanic service applicant ASVAB data, supplied by the Air Force Human Resources Laboratory will be used to estimate reliability, correct for range restriction, and assess practice effects.

RESULTS AND DISCUSSION

DESCRIPTIVE STATISTICS - Six subjects with post-AFQT percentile scores averaging 02, indicating extreme lack of motivation, were deleted from further analyses. This left a sample of 199. Computed means ($\bar{x}$) and standard

Table 1

Comparison of Means and Standard Deviations of Pre and Posttraining ASVAB Measures

Measure		Pre	Post	t
AR	Mean	41.9	42.6	1.7
	SD	4.8	6.0	3.3**
WK	Mean	37.5	41.1	8.6**
	SD	4.5	5.0	1.5
PC	Mean	38.0	43.3	7.7**
	SD	6.9	7.0	0.2
NO	Mean	50.9	53.7	4.4*
	SD	8.8	9.2	0.8
AFQT (Centile)	Mean	19.3	27.4	11.1**
	SD	5.8	10.9	10.6**

*Difference mostly due to practice effect.

**p less than .01, hypothesis that post is greater than pre.

Table 2

Sample Statistics for the Hispanic Enlistment Screening Test (HEST)

Measure	Mean	SD	Number of Items	Percent Correct
AR	27.9	5.2	35	80
WK	30.9	6.8	40	77
PC	13.2	3.4	20	66
NO	43.0	6.6	50	NA*
Total	115.0	14.2	145	79

*Not appropriate for speeded (NO) measure.

deviations (SD) for all variables are in Table 1. Subtests of AFQT are expressed in standard (T-score) metric with population mean of 50 and SD of 10. Initial ASVAB scores were low on all subtests except NO ($\bar{x}$ = 50.9). The lowest scores were on the most verbal of the subtests, WK and PC. These means were about 38, which corresponds to less than the 15th percentile. Pre-ASVAB SDs were considerably restricted due to the limited range (11-51 percentile) of AFQT in the sample. The mean AFQT percentile was 19.3. The raw score SD was 6.9. This is contrasted to an SD of 15.5 for a group of over 1,000 Hispanic service applicants (data from AF Human Resources Lab, 1984).

Retest scores were compared to pretraining scores via a test of correlated means (Guilford, 1965). Significant increases were found on the verbal subtests, WK and PC. Performance on PC was over five T-score points higher on retest. This indicates the AELT program had a measurable effect. The apparent increase on NO scores is misleading due to practice effects. Several studies (Sims & Maier, 1983, and unpublished study on Army applicants, 1982) have detected appreciable (about 2 T-score points) retest increases on NO but no increases on verbal subtests. When practice is statistically controlled, pre-post differences on NO are insignificant. There is no reason to expect AELT experiences to affect scores of nonverbal subtests such as NO. Differences between pre and postscore SDs were tested using a formula for correlated variances (Guilford, 1965). The only significant differences were increases on AR and AFQT.

Means for HEST subtests (Table 2) were high relative to the numbers of items. Percentages correct ranged from 66 on PC to 80 on AR. This indicates the subjects were generally able to comprehend and reason in written Spanish. Internal consistency reliability estimates were computed via Kuder-Richardson Formula 21. These estimates are shown in table 3 along with retest reliability data on like-named ASVAB measures for Hispanic applicants. The coefficients for corresponding HEST and ASVAB subtests are of comparable magnitude and are above .8 for AR and WK.

CORRELATIONAL ANALYSES - Correlations (r) of like-named HEST and pretraining ASVAB scores with posttraining ASVAB scores are given in Table 4. Low to moderate rs between pre and post-ASVAB scores were obtained. This was expected considering the range restrictions and low pretraining English verbal achievement in the sample. Correlations between similar pre and postsubtests were lowest for WK (.26) and PC (.15). In order to get an unrestricted population estimate of these rs, they were corrected for selection on AFQT (Guilford, 1965) using the raw score SD of 15.5 for Hispanic service applicants. Corrected rs for verbal scores were only .45 on WK and .30 on PC. Both uncorrected and corrected rs between HEST and corresponding posttraining ASVAB scores were higher than the rs involving pre-ASVAB on all subtests except NO. Only the comparison test (Guilford, 1965) for AR rs with postscores was statistically significant ($p < .05$), however.

Uncorrected and corrected rs with post-AFQT scores were also higher for HEST AR, WK, and PC than for like-named pre-ASVAB subtests. Comparison tests for both WK (HEST r = .49 vs. .38 for pre) and PC (HEST r = .46 vs. .31 for pre) were significant.

Table 3

Reliability Estimates for Subtests and AFQT Total

	ASVAB* (Retest)	HEST** (KR-21)
AR	.82	.81
WK	.85	.89
PC	.67	.64
NO	.71	NA
Total	.90	.89

*Based on Hispanic applicants for military service in 1983.
**Kuder-Richardson Formula 21 (Guilford, 1965).

Table 4

Correlations of Pre-ASVAB and HEST Subtests with Post-ASVAB Subtests and AFQT

Variables	ASVAB Pre - Post		ASVAB HEST - Post		
	r	r_c	r	r_c	t
AR vs. AR	.41	.62	.60	.70	2.0*
WK vs. WK	.26	.45	.39	.51	1.1
PC vs. PC	.15	.30	.25	.35	0.6
NO vs. NO	.46	.55	.32	.38	**
	ASVAB AFQT Pre - Post		**AFQT HEST - Post**		
	r	r_c	r	r_c	t
AR vs. AFQT	.27	.56	.44	.60	0.9
WK vs. AFQT	.03	.38	.30	.49	1.9*
PC vs. AFQT	.07	.31	.34	.46	2.5*
NO vs. AFQT	.24	.51	.23	.34	**

*p less than .05, hypothesis that HEST corrected r (r_c) is greater than pre r_c.
**No difference hypothesized.

Multiple correlations (R) with post-ASVAB scores as criteria were computed using corrected intercorrelations. Table 5 shows results of tests for increases in R squared (Bottenberg & Ward, 1963) when different combinations of predictors are included. The best combination for predicting post-AFQT scores consisted of pre-AFQT, and HEST AR and PC (R = .71). With only subtests as predictors, the optimal set included pre-ASVAB AR and NO, and HEST AR and WK. Among HEST subtests, NO did not add appreciably to prediction of post-AFQT. Due to their high intercorrelation (.68), HEST WK and PC were virtually interchangeable as predictors.

Multiple Rs in predicting post-ASVAB subtest scores from like-named pre-ASVAB and HEST subtests are also given in Table 5. As hypothesized, each HEST subtest made a significant contribution to prediction of the corresponding ASVAB subtest. The smallest increase involved HEST NO. As expected, the verbal subtests were the least predictable (WK R = .55 , PC R = .41).

CONCLUSIONS AND RECOMMENDATIONS - If Hispanics with low English skills are given language training which improves AFQT scores, then the HEST appears to make a significant contribution in predicting posttraining AFQT performance. This contribution also seems to be of practical importance. The amount of AFQT variance uniquely accounted for by HEST is about 10 percent, which is an increase of 25 percent over the variance explained by pretraining AFQT scores (50 percent vs. 40 percent). The significance levels reported should be interpreted with caution because corrected correlations have inflated standard errors (Bobko & Rieck, 1980). This does not affect the magnitude of rs, however.

The HEST NO subtest did not add to the prediction function of AFQT scores. Possible reasons for this include: (a) the nonverbal content of NO, (b) the pre-ASVAB NO correlated .51 with post-AFQT, thus not leaving much margin for improvement, and (c) timing problems with this brief, speeded subtest.

Data are being collected on several hundred more Puerto Rican recruits. In addition to seeing if the results of this study replicate, optimal subtest weights, and cutoff scores on composites will be sought. Possible group administration effects, particularly on NO will be investigated. Unless its contribution dramatically increases, NO will be recommended for deletion from the HEST. Before recommending it for operational use, the HEST should be normed on a female and male group which includes Chicanos and Cuban-Americans. Given satisfactory validation of the HEST, it might be administered at selected processing stations to identify nonqualifying Hispanics who would likely attain acceptable AFQT scores after remedial English training.

Table 5

Results of Multiple Correlation (R) Comparisons
Using Corrected Intercorrelations

Step	Predictors	Criterion	R^2	R	F
3	AFQT pre, ARH, PCH	AFQT post	.50	.71	5.8* (3 vs. 2)
2	AFQT pre, ARH	AFQT post	.48	.69	20.3** (2 vs. 1)
1	AFQT pre	AFQT post	.40	.63	-
4	ARH, NO pre, WKH, AR pre	AFQT post	.46	.68	5.3* (4 vs. 3)
3	ARH, NO pre, WKH	AFQT post	.44	.67	9.7** (3 vs. 2)
2	ARH, NO pre	AFQT post	.40	.63	12.2** (2 vs. 1)
1	ARH	AFQT post	.36	.60	-
3	ARH, WKH, PCH	AFQT post	.41	.64	2.6 (3 vs. 2)
2	ARH, WKH	AFQT post	.40	.63	9.9** (2 vs. 1)
1	ARH	AFQT post	.36	.60	-
2	AR pre, ARH	AR post	.53	.73	38.8** (2 vs. 1)
1	AR pre	AR post	.38	.62	-
2	WK pre, ARH	WK post	.30	.55	20.9** (2 vs. 1)
1	WK pre	WK post	.20	.45	-
2	PC pre, PCH	PC post	.17	.41	13.7** (2 vs. 1)
1		PC post	.10	.32	-
2	NO pre, NOH	NO post	.32	.57	6.0** (2 vs. 1)
1	NO pre	NO post	.30	.55	-

*p less than .05, hypothesis of increase in R^2.
**p less than .01, hypothesis of increase in R^2.

REFERENCES

Bobko, P. & Rieck, A. (1980). Large sample estimates for standard errors of functions of correlation coefficients. Applied Psychological Measurement, 4, 385-398.

Bottenberg, R. & Ward, J. (1963). Applied Multiple Linear Regression (PRL-TDR-63-6). Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division.

Guilford, J. P. (1965). Fundamental Statistics in Psychology and Education (4th ed.). New York: McGraw-Hill.

Sims, W. & Maier, M. (1983). The Appropriateness for Military Applications of the ASVAB Subtests and Score Scale in the New 1980 Reference Population (Memorandum 83-3102). Alexandria, VA: Center for Naval Analyses.

Attrition and Performance Ratings of ESL Soldiers in BT*

Harvey Rosenbaum

American Institutes for Research
Washington, D.C.

The Army's Basic Skills Education Program (BSEP) is an on-duty, job related, basic skills development program in the areas of literacy, math, and English-as-a-second language (ESL). One of BSEP's goals is to improve the ability of soldiers to perform satisfactorily in training and on the job. BSEP I programs are intended to provide soldiers with the basic competencies to complete Initial Entry Training (IET). BSEP II programs are to provide soldiers, E-1 through E-5, with the job related competencies necessary for their military duties. A fundamental notion underlying the BSEP program, for which there is some support (Krug, et al., 1984), is that soldiers with basic skill deficiencies are less likely to perform their job well, are less successful in the Army, and display a higher rate of attrition.

This paper presents additional support for the notion motivating BSEP using attrition data and performance rating data for a group of soldiers who completed the BSEP I ESL Program. The data show that level of English language proficiency is linearly related to attrition in Basic Training (BT) and tends to be related to drill sergeants' ratings of soldiers' BT performance. The attrition data also provide a basis for estimating the percent of attrition reduction in BT due to the ESL program.

BACKGROUND

In the summer of 1982, the Army implemented a new six-week BSEP I ESL Program at eight TRADOC installations. The new program is functionally, or BT, oriented and intended to provide soldiers with information necessary for success in BT as well as improve their English language skills. The BT information was drawn from 25 tasks included in the Soldier's Manual of Army Testing or SMART book.

Between September 1983 and May 1984, the American Institutes for Research (AIR) conducted an evaluation of the new ESL program using a variety of program and follow-up measures (Rosenbaum and Stoddart, 1984). One of the pre/post measures used to assess the effect of the program on language proficiency was the English Comprehension Level Test (ECLT). The ECLT, developed by the

*This research was conducted for the U.S. Army Research Institute, Contract Nos. MDA-903-81-C-AA04 and MDA-903-84-C-0128. This paper does not necessarily express the official opinions or policies of the contracting agency.

Defense Language Institute English Language Center (DLIELC), is probably the most widely used English language test in the Armed Services. It is a standardized, group administered test of English comprehension with a strong reading component. Data gathered in previous studies by AIR (Holland, et al., 1982; Rosenbaum, 1983), indicate that as a measure of group performance, the ECLT is an effective indicator of oral comprehension and speaking proficiency.

Eligibility for the ESL program is largely determined by the soldiers' ECLT score. Soldiers with ECLT scores below 70 are automatically eligible for the program.

Included in the follow-up measures used to determine program effects, were data on BT attrition and performance ratings completed by drill instructors on an available sample of soldiers in BT. Data presented in the following sections were gathered as part of AIR's overall evaluation of the ESL program.

METHOD

Subjects

ECLT scores and BT attrition data were available for 582 soldiers. Supervisors' ratings of soldiers' performance in BT were also obtained for an available sample of 156 soldiers who had completed the program. Nearly 90 percent of the soldiers were native speakers of Spanish and more than 70 percent of these were from Puerto Rico. Approximately 75 percent of the sample graduated from high school and nearly 25 percent graduated from two- or four-year colleges.

Procedure

Soldiers who appear potentially eligible for the ESL program are usually administered the ECLT when they arrive at their training base. Those with ECLT scores below 70 are usually placed in the program within a week or two of their ECLT testing. Soldiers are retested with the ECLT when they complete the program.

Installation education centers provided AIR with information on whether their students completed BT or were discharged from BT. Most education centers, however, do not routinely receive this information from the BT units and must make special efforts to obtain it. One education center did not provide any data on attrition and most of the other centers provided attrition data on only part of their student enrollment. Consequently, we were only able to obtain attrition data on a portion of the nearly 1800 soldiers involved in the overall evaluation.

Performance ratings were obtained for an available sample of soldiers during their last four weeks of BT. AIR researchers had drill sergeants rate soldiers' performance on 14 BT activities by

comparing their performance with all other soldiers. All activities required the use of language and were taken from the Program of Instruction for BT. The descriptions of the activities were pilot tested with drill sergeants at Fort Dix to insure that they were understandable and relevant. The rating system was a four-point scale with the values of better than most, as well as most, not as well as most but gets by, and performs inadequately. Examples of the BT activities are "responds correctly to questions," "uses challenge and password" and "reads SOP for inspections."

RESULTS

Attrition

Soldiers who were discharged from the Army during BT displayed poorer performance on the ECLT than those who completed BT as shown in Table 1. Discharged trainees had lower entry ECLT scores, lower exit ECLT scores, and made smaller ECLT gains.

Table 1
Comparison of Mean ECLT Scores

Attrition category	Entry ECLT	Exit ECLT	ECLT gain
Completed BT	42.9	57.6	14.8
(n)	(505)	(507)	(505)
Discharged from BT	37.4	46.3	9.8
(n)	(78)	(75)	(75)

The attrition rate is linearly related to exit ECLT scores with soldiers scoring below 30 having an attrition rate more than five times that of soldiers scoring above 69 as shown in Table 2.

Table 2
Comparison of Soldiers Completing BT with Those Who Were Discharged from BT

	Exit ECLT score						
Attrition Category	0-29	30-39	40-49	50-59	60-69	>69	n
Completed BT	64%	75%	84%	91%	92%	93%	507
Discharged from BT	36%	25%	16%	9%	8%	7%	75
n	36	63	95	123	119	146	582

Performance Ratings

Sergeants tended to assign a soldier similar performance ratings for most of the 14 activities. Two sets of Pearson Product Moment correlations were conducted to determine the degree of similarity between the activity ratings. In order to conduct these correlations, ratings were given numerical or score values by assigning a value of one through four to the activity ratings: the value of one was assigned to performs inadequately, two was assigned to not as well as most, etc.

Soldiers' performance ratings for each activity were then correlated with each of the remaining 13 activities yielding 91 statistically significant correlations between .51 and .81. For the second set of correlations, each activity rating was correlated with the average of all 14 ratings for each soldier yielding a correlation range of .77 to .86.

The strong correlation of the 14 activity ratings justifies using the soldiers' average ratings as unitary performance ratings. To relate soldiers' average performance ratings to their exit ECLT scores, soldiers were divided into two categories: average performance ratings of 2.4 or less were assigned to the performs not as well as most category, and average ratings of 2.5 or more were assigned to the as well as most or better category.

Soldiers with lower exit ECLT scores tended to receive lower performance ratings. For example, soldiers with ECLT scores below 60 are nearly three times as likely to be in the performs not as well as most category than are soldiers with higher ECLT scores as can be inferred from Table 3. However, the distribution of soldiers in the category performs not as well as most by exit ECLT score is not clearly linear. Two factors that may be affecting these data are (1) the sample size below 50 ECLT is small, and (2) BT performance ratings may be less directly related to language proficiency. During interviews with AIR researchers, drill sergeants often stated that a soldiers' attitude and motivation are major factors in completing training.

Table 3
Comparison of Soldiers Rated as Performing Not as Well as Most With Soldiers Receiving Higher Ratings

	Exit ECLT						
Performance Rating Category	0-29	30-39	40-49	50-59	60-69	>69	n
As well as most or better	67%	50%	79%	69%	89%	90%	124
Not as well as most	33%	50%	21%	31%	11%	10%	32
n	9	12	19	32	36	48	156

DISCUSSION

Level of English proficiency is clearly a factor in soldiers' success in BT. The data indicate that limited proficiency in English contributes to attrition and poorer performance. Because it increases soldiers' proficiency in English, as measured by the ECLT, the BSEP I ESL program can be assumed to be reducing attrition and improving soldiers' performance in BT. The sample of nearly 1800 soldiers that completed the ESL program during the evaluation period showed a mean ECLT gain of 15.0. This mean gain is as good or better than the mean ECLT gains obtained from previous Army ESL programs.

The reduction in BT attrition attributable to the ESL program can be estimated by comparing the attrition of soldiers completing the ESL program with a hypothetical group of eligible soldiers receiving BT without the benefit of the program. Computing this estimate requires one conservative assumption in addition to the available data. The assumption is that the attrition rate at each exit ECLT level is also applicable for soldiers who have not participated in the ESL program. Even if incorrect, this assumption is most likely to result in an underestimate of the reduction in attrition due to the ESL program since soldiers in the attrition sample are different from a hypothetical group of new trainees with matching English proficiencies. The sample group started BT already knowing a great deal of technical information and should find BT easier than the hypothetical group. Interviews with soldiers during the BT follow-up phase of the evaluation support this view. Many soldiers stated that BT was easy because of the information they had learned in the program, others claimed that they would not be able to complete BT without the program.

We can estimate the BT attrition for 1000 eligible soldiers who were not enrolled in the ESL program and compare this with the attrition estimate for 1000 soldiers enrolled in the program. The process of estimating the attrition for 1000 not enrolled soldiers is presented in Table 4. Line one shows the pre-program distribution of ECLT scores for the sample of 1762 soldiers used in the ESL program evaluation. The next line indicates the number of soldiers per thousand at each ECLT range. The third line provides an index of attrition for each ECLT level based on the percent of attrition for that level as given in Table 2. The final line in Table 4 indicates the number of soldiers estimated to attrite at each ECLT level yielding a total of 185 per 1000.

Table 4
Procedure for Estimating BT Attrition of 1000 Eligible Soldiers Not Enrolled in ESL Program

	ECLT Levels					
	0-29	30-39	40-49	50-59	60-69	Total
Distribution of SM by entry ECLT	18%	22%	18%	21%	21%	100%
No. of SM/1000	180	220	180	210	210	1000
Index of attrition	.36	.25	.16	.09	.08	- -
No. of SM attriting	65	55	29	19	17	185

Using the same procedures, Table 5 shows the basis for the attrition estimate for 1000 soldiers enrolled in the program yielding an attrition total of 116. According to these estimates, the Army's ESL program saves 69 out of every 1000 soldiers from attrition in BT or reduces BT attrition by 37 percent.

Table 5
Procedure for Estimating BT Attrition of 1000 Soldiers Enrolled in ESL Program

	ECLT Levels						
	0-29	30-39	40-49	50-59	60-69	>69	Total
Distribution of SM by exit ECLT	3%	9%	15%	20%	26%	27%	100%
No. of SM/1000	30	90	150	200	260	270	1000
Index of attrition	.36	.25	.16	.09	.08	.07	1.00
No. of SM attriting	11	23	24	18	21	19	116

REFERENCES

Holland, V.M., Rosenbaum, H., Stoddart, S.C., & Redish, J.C. BSEP I/ESL programs - Volume one: Findings. Washington, D.C.: American Institutes for Research, October 1982.

Krug, R.E., Hahn, C.P., & Wise, L.L. Review of BSEP I and BSEP II programs. Task One Report. Washington, D.C.: American Institutes for Research, July 1984.

Rosenbaum, H., Hahn, C.P., and Holland, V.M. Testing the English language proficiency of soldiers who are not native English speakers. Paper presented at the 25th Annual Conference of the Military Testing Association, Gulf Shores: Alabama, October 1983.

Rosenbaum, H., & Stoddart, S.C. Evaluation of the Functional Pre-BT ESL Course. Washington, D.C.: American Institutes for Research, December 1984.

90S

STANDARD SETTING METHODS

FOR SKILL QUALIFICATION TESTS (SQTs)

Dr. Allan L. Pettie

United States Army Training Support Center, Fort Eustis, Virginia

The purpose of this paper is to examine three different methods for setting passing scores for Skill Qualification Tests (SQTs). The Skill Qualification Testing Program is one component of the United States Army's Individual Training Evaluation Program. The other components, the Common Task Test and the Commander's Evaluation, are hands-on tests. The SQT component is a task-based, performance-oriented objective test. Enlisted soldiers are tested by Military Occupation Speciality and Skill Level.

Since the implementation of SQTs, the Army has used the same passing score of 60 for all SQTs. Experience has shown that the adoption of this common standard, which failed to consider test difficulty and content, and its application across the broad spectrum of diverse MOSs was unjustified. Starting October 1, 1986, this policy of a common standard for all SQTs will be discontinued. A separate passing score, based upon information about that SQT, will be set for each SQT. Prior to the implementation of this change, pilot studies of three methods, which were deemed applicable to the Army, were conducted. These three methods (untrained examinee, performer non-performer, and validation data) are described briefly in the following sections.

UNTRAINED EXAMINEE METHOD

The rationale for the setting of a passing score by this method is that the trained soldier should be able to do at least as well as the atypical untrained examinee. Factors, not related to competence, but which allow the incompetent to pass, such as test wiseness, the effects of tests which measure primarily general aptitude, reading ability, and common sense could be minimized by this method. Scores of untrained examinees could be indicative of the extent to which the SQT measures general aptitude and reading ability.

The untrained examinees of choice were soldiers in the same or related MOS who were in the first week of Advanced Individual Training. It was presumed that these examinees would be similar in background and representative of the trained group before training. After the SQT was administered, the passing score was set at the eightieth percentile. A soldier who scored above the eightieth percentile was considered to be atypical of the untrained group; thus, minimum competence is defined on the basis of the untrained examinee.

909

PERFORMER NON-PERFORMER METHOD

This adaptation of the contrasting group method (Livingston and Zieky, 1982) relies upon the judgment of supervisors to form a performer group and a non-performer group. Supervisors were instructed to base their judgments of MOS competence on criteria other than test scores. Costs of classification errors, performers who fail and non-performers who pass, were considered to be equally serious. The passing score was determined to be the point of intersection for the performer and non-performer distributions.

VALIDATION DATA METHOD

Out of a desire to take advantage of existing soldier tryout data, this method was formulated. Soldier tryout data are normally collected at the task level, a mini-test of three or more questions. Soldiers, through self-ratings or supervisory ratings, are designated performers or non-performers on the particular task. Since the data for a performer group, consisting of the same performers, and a non-performer group, consisting of the same non-performers, do not exist, a composite performer group and a composite non-performer group were formed.

The performer group, actually task-level performers, thus resulted in testees who were performers on all tasks. Similarly, the non-performer resulted in a testee group which was made of non-performers on all tasks. Variances were calculated at the task-level and then summed across tasks. Since the scoring for SQTs is task-based, an estimate of the standard deviation was calculated by an adaptation of standard deviation estimation by items (Tinkleman, 1971).

Similar to the performer non-performer method, costs were equalized, and the passing score was set at the point of intersection for the performer and non-performer distributions.

RESULTS

Since the pilot studies were conducted by the untrained examinee and validation data methods or the performer non-performer and validation data methods, passing scores set by the method pairs are presented in Table 1.

TABLE 1

PASSING SCORES FOR UNTRAINED
EXAMINEE, PERFORMER NON-PERFORMER
AND VALIDATION DATA METHODS

SQT	UNTRAINED EXAMINEE	VALIDATION DATA	PERFORMER NON-PERFORMER
15B1	36	61	
16P1		48	72
16R4		63	63
24G1		70	70
24G3		67	67
24A1	50	74	
24H3	50	76	
27E1		61	53
27E3		63	64
31V1	55	63	
31V2	58	65	
31V3	50	65	
31V4	49	67	
42E3	38	60	
43E1	42	68	
43E3	47	64	
64C1	73	63	
64C3	81	61	
71Q3	41	56	
71Q4	41	53	
71R1	55	67	
71R2	43	67	
71R3	47	61	
71R4	47	61	
73D1	64	62	
73P3	62	64	
75B2	40	61	
75B3	31	67	
76P1	39	61	
76P3	43	57	

Passing scores set by the untrained examinee method demonstrate greater range than the other two methods. The 75B2 SQT has the lowest passing score of 31, and the 64C1 SQT has the highest passing score of 81. Passing scores for the validation data method ranged from a low of 48 for 16P1 SQT to a high of 76 for the 24H3 SQT. Passing scores for the performer non-performer method ranged from a low of 53 for the 27E1 SQT to a high of 72 for the 16P1 SQT.

Comparisons of the untrained examinee method to the validation data method show that except for four SQTs (64C1, 64C3, 73D1, and 73D3) passing scores for the untrained examinee method are lower than the validation data passing scores. Passing scores set by the validation data and the performer non-performer methods, with the exception of 16P1 and 27E1, are very close in value. Upon analysis of the performer non-performer data, a possible explanation for the discrepancy was discernible. This data consisted of seven cases, four 25's and three 75's, for the non-performer group. Perhaps, the three soldiers who scored 75 were misclassified. In any event, the non-performer group appeared to be widely disparate in ability. No expalantion was discovered for the 27E1 discrepancy.

CONCLUSIONS

There was no substantial agreement or pattern of disagreement between the passing scores set by the untrained examinee and the validation data methods. Untrained examinee passing scores were generally lower and considerably different from the validation data passing scores. The discrepancy between the two estimates ranged from a difference of two points (73D1 and 73D3) to a 36 score point difference (75B3). Given the different rationales for the two approaches, this lack of agreement should be expected. The crucial question is which rationale is best applicable to the Army's needs.

Better agreement was reached for the validation data and the performer non-performer methods. For four of six SQTs no difference was observed between the passing scores. Considerable differences were noted for two SQTs. A possible explanation for the 16P1 difference was noted above, but the 27E1 difference of eight score points is troublesome.

The validation data method was developed as a substitute for the performer non-performer method. It appears that the validation data method may be a viable alternative to the more resource intensive performer non-performer method. The validation data method is less resource intensive, because of the collection of the data does not represent a new requirement. The

validation data method is attractive, because of the existing requirement for the collection and maintenance of the data and because it represents no new requirement upon the test developers.

Examination of the setting of passing scores will continue along three different lines. Few pilot studies were performed for the performer non-performer method; therefore, further analysis with more SQTs needs to be conducted to examine the viability of the validation data method. Another form of analysis will be the application of multiple matrix sampling estimation procedures to estimate the mean and standard deviation for the performer non-performer groups from soldier tryout data. Although the random assignment of examinees to items assumption of this method will be violated, the application of this method should be examined. Finally, the proximity of the passing score to an acceptable level of competence and the reasonableness of passing scores set by the untrained examinee method will be examined.

References

Livingston, Samuel A., and Zieky, Michael. Passing Scores. Princeton, N. J.: Educational Testing Service, 1982.

Tinkleman, Sherman N., Planning the Objective Test. In R. L. Thorndike (ed.). Educational Measurement. Washington, D. C., Americal Council and Education, 1971, pp 65-66.

Assessing Tank Commander and Gunner Proficiency on U-COFT

Scott E. Graham
and
John A. Boldovici

Army Research Institute-Fort Knox Field Unit

The Armor community is striving to improve its selection and training of M1 Tank Commanders (TC) and gunners. If these goals are to be accomplished, valid performance measures must be established which assess the full range of tank crewmen duties. One of the most pressing needs in Army evaluation is the development of tests whose psychometric properties are known and which inspire confidence (Boldovici & Sabat, 1985).

The reliable measurement of hard, combat-oriented skills and soft, leadership skills has not been successfully accomplished. The lack of standardized and stable performance measures makes it difficult to compare individual and crew performance across time and units. As a result, NCO promotions may be yoked to Time in Grade rather than be performance based. Inconsistencies in training evaluation research have also resulted from the lack of valid performance criteria. A number of criterion measures have been developed for assessing gunnery-related performance. Continuing credence is, however, placed on scores obtained from live-fire gunnery exercises, and in particular, Table VIII. Unfortunately, problems affecting the reliability of these scores, e.g., varying weather and equipment conditions, make questionable the comparison of scores across days, ranges, and units.

Training simulators and other electronic training devices provide new opportunities for evaluating the performance of Armor crewmen. Their main advantages include precise presentation of target conditions with accurate scoring and timing. Graham and Black (1985) have identified a number of critical TC tasks which can be evaluated on the newly developed M1 Unit-Conduct of Fire Trainer (U-COFT). These target engagement tasks include laying the main gun, issuing fire commands, boresighting and degraded mode gunnery.

The research described here develops and evaluates a U-COFT proficiency test for M1 gunners. The primary goal is to assess the psychometric characteristics of the U-COFT as a testing device and to evaluate the feasibility of testing gunners alone, i.e., independent of TC performance.

Method

U-COFT Test Development. The U-COFT Gunner's Test contained eight shortened exercises, each with four engagements, selected from the U-COFT's TC/Gunner's training matrix. One target was friendly (an M2) making a total of 31 target engagements. The engagements were selected to match target conditions found in Table VIII of the M1 tank combat tables. For example, half of the engagements in each were ownvehicle stationary (or moving), and half were long (short) range. Roughly equivalent numbers of single/multiple and stationary/moving targets were included. Table VIII has 20% battlesight engagements (as opposed to precision or "full up") while the U-COFT test has 25%. Half of the engagements also required thermal sights.

Scoring. A number of performance measures were obtained from each engagement. These included Hitrate which was defined as the number of hits divided by the number of targets presented. Other measures included First Round Hitrate, Azimuth and Elevation errors, Target Identification (ID) time,

which was the time from when the target appeared until the gunner said "identified," and Opening time, which was the time from target appearance until the first round was fired.

The U-COFT software package reports three composite performance scores which were also recorded. The Target Acquisition score measures target acquisition time and identification/classification errors (U-COFT Handbook, 1985). The Reticle Aim score assesses opening time, time to kill, and reticle aim error. Lastly, the System Management score counts pre-firing switch errors, ammunition errors, and excessive ownvehicle exposure times. Each of the scores is reported as a letter grade, A, B, C, or F with corresponding numerical values of 4.0, 3.0, 2.0, and 1.0.

Participants. The U-COFT gunner's test was administered at the completion of other research evaluating training transfer between the U-COFT and the MK-1 videodisc gunnery simulator (Witmer, in prep). The 32 soldiers used were M60A3 loaders and drivers from the 194th Armor Brigade at Ft Knox, KY. The majority had ranks of Private First Class. The soldiers, with few exceptions, had not served as gunners other than in Advanced Individual Training, and had no experience with the M1 tank.

Three TCs were predominately used with a fourth TC used for one session. One civilian TC was an ex-General Electric employee who had hundreds of hours of U-COFT experience as a U-COFT TC (COFT-experience). The other two were a Sergeant First Class from the Armor School Weapons Department (Sr NCO) and a Sergeant from the 2/6 Cavalry (Jr NCO), a training support unit of the Armor School. They had no prior U-COFT experience.

Procedure. The gunners received 1 1/2 hours of U-COFT training during the U-COFT/MK-1 transfer study. The U-COFT gunner's test began with eight practice engagements. The last four required use of the Gunner's Auxiliary Sight (GAS) as there was simulated failure of the laser rangefinder, stabilization system, Gunner's Primary Sight and computer system. The TC trained the gunner on use of the GAS and how to fire with manual lead and elevation.

The eight subtests were sequentially presented with a short pause between each. During this time, the U-COFT Instructor/Operator (I/O) had to terminate the standard 10 engagement exercise, dump the printouts, and enter the six-digit code for the next subtest. This procedure was awkward and a few printouts were missed. Following a break, a retest was presented, which consisted of a different subtest order. The practice, test, and retest took approximately 2 hours.

An attempt was made to minimize the effects of differential TC performance by having the I/O talk the TC onto the target. The I/O might have said, for example, "next target, a T-72, is left in 10 seconds." This modification theoretically reduced the variability of target identification times across TCs and minimized fire command errors. The standard U-COFT procedure also requires the gunner on defensive engagements to move his head out of the GPS, check the GAS to see that the gun has cleared the berm, say "driver stop", and then go back to the GPS. This procedure was omitted.

Results and Discussion

For each of the dependent variables, test-retest reliability coefficients (Pearson r) were computed. The reliability coefficient for Hitrate (.80), First Round Hitrate (.72), Target ID Time (.87), Opening Time (.22), and the U-COFT Target Acquisition (.76) and reticle aim (.83) scores are encouraging.

Demonstrated reliability is a necessary component for valid tests and these U-COFT measures appear appropriate. The next logical step is to validate the measures against known and acceptable performance criteria, i.e., combat effectiveness measures. The typical inclination is to validate device-mediated gunnery tests against live-fire gunnery scores. Unfortunately, the poor psychometric properties of the live-fire measures virtually assure a weak relationship at best. The validation and acceptance of U-COFT tests will likely result from its convergence with various gunnery performance variables, sound military judgement, and face validity.

The data show poor reliability for Azimuth (.42) and Elevation (-.07) errors and the U-COFT System Management scores. The elevation and azimuth error unreliability appears to result from a tight distribution with some extreme scores. The mean elevation error of .57 mils is within the hit range of most targets, yet some individual subtest elevation errors exceeded 7.0. The U-COFT Reticle Aim score reduces the effects of extreme errors by scoring an "F" for any missed target. While the score does reflect time in addition to lay error, the reliability for Reticle Aim score is .83. This suggests Reticle Aim could be a reliable measure.

The unreliability of System Management errors (.11) may be artifactual of the U-COFT procedure. The gunners were instructed to leave the gun select switch on "main gun". On defensive engagements when the crew was "killed" as the result of being exposed too long, the gun select switch automatically reset to safe. Random switching errors may have then resulted from failure to put the switch back on main gun. Other system management errors resulted from incorrect ammo select switch settings.

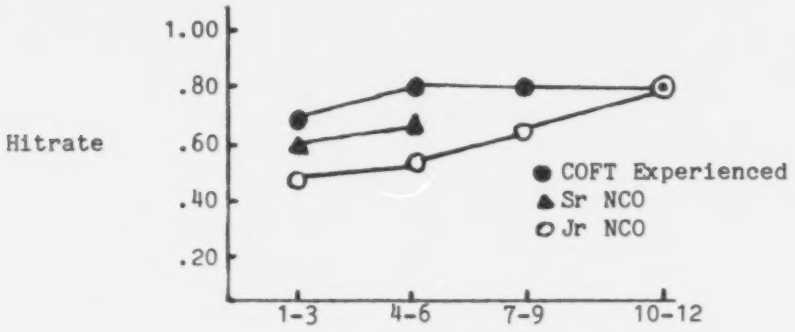
The feasibility of using U-COFT as a testing device depends, in part, on its ease of the administration and scoring. Three U-COFT scores, Reticle Aim, Target Acquisition, and System Management, could well be used as performance measures. The composite mean of these scores yielded a test-retest reliability of .82, and correlated .87 with Hitrate. These data substantiate the internal validity of the U-COFT scoring procedure.

Performance on the eight subtests varied considerably. Mean Hitrate for single long range stationary targets at night was .91, while the hitrate for multiple short range moving targets in degraded mode was only .19. Multiple and moving targets, not surprisingly, were most difficult, although the hitrate for short range single moving targets was .83.

One minor difficulty with the U-COFT for training and testing is that it includes dispersion rounds. A gunner may have a perfect sight picture, fire within the required time limits, and still miss the target. The opposite also occurred, although less frequently. While the dispersion rounds were likely included to match live ammunition characteristics, they result in greater unreliability of testing scores, and random bad feedback for trainees. Software updates should include the capability of turning off the dispersion rounds.

TC Performance Effects. Despite attempts to minimize TC effects, U-COFT performance differed as a function of TCs, i.e., COFT-experienced, Sr NCO, and Jr NCO. Gunners paired with the COFT-experienced TC had a combined test and retest Hitrate of 74%, while those paired with the Sr NCO shot 64%, and those with the Jr NCO shot 63%, $F(2,28) = 3.89$, $p < .05$. Similar significant differences for TCs were found for First Round Hitrate, ID time, and the Reticle Aim, Target Acquisition, and System Management scores.

The data were recomputed to reflect these changes over sessions. The figure below shows changes in Hitrates for the three TCs over the duration of the experiment. Each TC session reflects the mean performance of three different gunners.



Hitrate is ostensibly a measure of gunner performance as the gunner makes the final lay of the gun and pulls the trigger. Hitrate is shown here to be also a function of TC performance. The first three gunners paired with the Jr NCO hit 48% of the targets while his last three gunners hit 77%. Each session gave the TC about 3 1/2 hours of U-COFT experience, including the U-COFT time in Witmer's experiment. Three TC sessions, therefore, correspond to about 10 hours of U-COFT time. The abscissa in the figure could alternatively be labled 10, 20, 30 and 40 hours of TC COFT time. These data show that the COFT-experienced TC's contribution to hit performance asymptoted between 10 and 20 hours, while the Jr NCO's contribution had not stabilized at 40 hours.

Several factors may be contributing to this pattern. First when target ID time is plotted as a function of TC sessions, improvement in ID time for the three TCs essentially mirrors Hitrate performance. Secondly, the TCs might becoming better trainers of the novice gunners over sessions. The TCs were learning what errors were typically being made by the gunners and were better instructing the later gunners in these areas, e.g., use of the GAS. While the observations are subjective and non-systematic, improvement in training was particularly noticeable for the Jr NCO. At first he nervously barked his fire commands, and paid little attention to the gunner. In later sessions, he readily assisted the I/O in the U-COFT orientation and praised the gunners during breaks. General Electric officials independently reported a similar phenomenon during their COFT tests. TCs and gunners were at first blaming each other for misses, but over time greater cooperation developed.

The results suggest at least three things. First, under the conditions used in this experiment, the U-COFT is not appropriate for testing gunner proficiency alone, as the gunner's performance was not independent of the TC. Confederate TCs might first be trained to stable performance levels, but this would require considerable time and resources, e.g., 40 plus hours. If a gunner alone test is desired, the soon-to-be-fielded Institutional-Conduct of Fire Trainer (I-COFT) may be more appropriate. The I-COFT will have essentially the same hardware as the U-COFT but software modifications will permit voice synthesized fire commands. The U-COFT has been developed as a TC/gunner training device, and for the most part, is better suited for testing the proficiency of TC/gunner pairs.

Second, the results reinforce the notion that crew performance is largely a function of the TC's ability to train his crew. More time might well be spent on training the TC how to train his crew.

Third, U-COFT performance results from a number of factors, in addition to the gunner's ability to quickly and accurately lay the gun. Some of the factors may be idiosyncratic to the COFT itself and not related to live-fire gunnery performance. The improvement of the Jr NCO over 40 hours of training may have resulted from a growing familiarity with the device and the computer-generated graphics, and not from changes in his gunnery ability. Had the TC been going through the U-COFT training matrix, he would likely have been half way through the matrix before his performance stabilized. Improvements in performance might be unrelated to the systematic progression of target conditions, but of general familiarity with the machine. The built-in training package is, nevertheless, better than most other training devices which have none. The validity of the U-COFT as a testing device is threatened by the fact it takes so long to get stable performance.

Modeling Gunnery Performance. The precise target presentation and response recording capabilities of the U-COFT result in accurate part-task gunnery data, which can be used in the development and refinement of combat models. Various attrition-rate models exist with factors such as time to acquire target and time to fire following a hit (or miss). The value of these complex models has been limited, however, by poor parameter estimates. The estimates have largely come from past battle data which are sketchy at best, as armies engaged in war are interested in a number of factors in addition to gathering quality data for combat models. The U-COFT has the capability of gathering reliable estimates under a number of conditions, including future battlefield conditions. As such, the U-COFT can be used to evaluate alternative future training strategies.

In the present research, stepwise regression analyses were performed to help understand what factors were underlying the various performance measures. Hitrate was found to be primarily determined by reticle aim accuracy, as target ID time and Opening time failed to significantly load into the equation. In the previous discussion of factors affecting changes in hitrate over TC sessions, changes in target ID time were shown to parallel changes in hitrate. The regression analysis suggests, however, that changes in target ID time are unrelated to changes in Hitrate, once the effects of aiming error are removed. This finding lends additional credibility to the explanation that improvement over TC sessions resulted from improved training of the gunner.

General Discussion

The results demonstrate the U-COFT holds great potential as a device for assessing gunnery proficiency, although not necessarily for gunner-alone tests as tried in this research. Test-retest reliability coefficients exceeding .80 were found for a number of performance measures including Hitrate and target ID time. These measures are well within the acceptable psychometric range for military tests. In addition, U-COFT tests might be used as criteria against which other less expensive or portable tests could be validated.

The U-COFT proficiency test developed for this research mirrored target conditions in the M1 tank Table VIII. This may not be necessary. As Boldovici (1979) noted, the lowest level of "enabling skills" required for

gunnery performance, e.g., psychomotor skills and system procedural knowledge, are highly redundant across engagement conditions. As a result, it is unlikely that a crew who is relatively good at long range moving targets at night would be poor at short range stationary targets in daylight.

The high Hitrate for single stationary targets further demonstrates the relative ease of using the M1 fire control system when it is fully operational. Considering that the experimental gunners were M60A3 loaders and drivers with little gunners' experience, a ceiling effect would be expected if the easier target engagements were used with experienced TC/gunner pairs. Perhaps the best feature of the U-COFT is its ability to train and test under various degraded conditions. Future U-COFT tests might place even greater emphasis on evaluating degraded gunnery performance, e.g., manual control conditions.

Improved evaluation of tank gunnery skills with the U-COFT can lead to a stronger Armor force. Skills and abilities other than those measured by the U-COFT are, however, equally important and should not be overlooked. Graham and Black (1985) found that soft skills, e.g., ability to train, ability to communicate, and general leadership qualities, were predominately identified as the distinguishing characteristics of TC excellence. Likewise, the results described here show the TC's ability to train his gunner has a large effect on gunnery performance. Continuing development of evaluation batteries which assess both hard combat-oriented skills and soft leadership skills is needed, if the force is truly to become an Army of excellence.

References

Boldovici, J.A. (1979). Analyzing Tank Gunnery Engagements for Simulator-Based Process Measurement. Army Research Institute Research Report 1227.

Boldovici, J.A. & Sabat, S.R. (1985). Measuring Transfer from Training Devices to Weapon Systems, Paper presented at NATO Symposium on the Transfer of Training to Operational Military Systems, Brussels, Belgium.

Graham, S.E. & Black, B.A. (1984). Defining and Assessing Tank Commander Excellence, Army Research Institute Research Report 1401.

Unit-Conduct of Fire Trainer Instructor's Utilization Handbook (1985). General Electric Company, Daytona Beach, FL.

Witmer, B.G. (in Prep). Unit-Conduct of Fire Trainer (U-COFT) and Videodisc Gunnery Simulator (VIGS) Cross-Training and Training Transfer. ARI-Ft Knox Field Unit Working Paper.

Criterion Referenced Testing for the U.S. Navy's Nuclear Submarine Fleet

Jeffrey A. Cantor and Lee Walker
DDL-OMNI Engineering
McLean, Virginia

Introduction

The design and development of criterion-referenced System Achievement Tests (CR SATs) for each of the Naval Enlisted Classifications (NECs) associated with the Fleet Ballistic Missile (FBM) Strategic Weapon System (SWS) Training Program is a goal of the U.S. Navy's Strategic System Program Office (SSPO). DDL-OMNI Engineering has been involved in the design, development and refinement of a method to accomplish this goal over the past two years. The methodology which will be described in this paper presentation has evolved as a result of the planning, trial and refinement on the part of both SSPO and DDL-OMNI Engineering.

Background

The CR SAT is a test which is administered to all SWS technicians periodically to measure overall system knowledge and skill levels based on the specified jobs/tasks and procedures identified as essential for each NEC. The primary purpose of the SAT is to evaluate the readiness of the personnel to operate the equipments within their NEC, so as to accomplish the ship's mission; its secondary purpose is to assess the knowledge and skill of the personnel tested. Previous efforts at a test design specification for a SAT had generally been in terms of the format and content of a previous version of the SAT within a given NEC. This procedure did not allow for control in terms of overall coverage of all requisite jobs/tasks and procedures, nor did it allow for updating of information coverage necessary as a result of changes in technology of the equipment and changes in procedures. The new CR SAT program was initiated to allow for such changes.

Test Design Specification

The basis for the construction of a CR SAT is the test design specification. This specification is an engineering blueprint of the knowledges identified as necessary to possess in order to effectively perform as a competent FBM/SWS technician in the specific NEC for which the CR SAT is written.
The project undertaken to produce a test design specification allows for inputs to ensure adequate content coverage and changes in technology and at the same time, be self-sustaining in a dynamic NAVY personnel system. The project team includes contractor subject-matter experts, training data analysts, and competent and knowledgeable NAVY technicians and supervisors representative of the NEC under development. The methodology devised for the test design specification is predicated on the use of a committee of technical experts with current knowledge of the jobs/tasks and procedures necessary for successful performance in the NEC for which the specification is to be constructed.

The contractor project leaders, in preparation for the test design specification workshop, prepare a comprehensive listing of those jobs/tasks and procedures identified as necessary to the function of the NEC. This listing of jobs/tasks and procedures comes from applicable Personnel Profile

Tables (PPPs), Training Level Assignments (TLAs), and other procedural documentation literature which supports the equipment systems operated by the respective NEC. These listings are then prepared in the categories of the applicable equipment system/subsystem or to the applicable area of administration/security, casualty operating procedures, or watch qualifications. The basic test will include approximately 360 question items, divided into sections of equipments (approximately 270 items), administration-/security (30 items), watch qualifications (30 items), and casualty operating procedures (30 items).

Test Design Workshop Process

The workshop takes the form of a five-day convening, with technician representatives from each of the Submarine Groups. Since two levels of tests will be constructed for each NEC, one at the watchstander level and one at the supervisor level, personnel representative of each of the levels are identified to participate. In actual operation, two concurrent workshop convenings are held. The technicians are first oriented to the purpose of the workshop. During the week-long workshop process, the technicians will be providing feedback regarding each of the above cited categories. In the case of the administration/security category, the technicians provide inputs via a Q-SORT process. In this Q-SORT, each technician prioritizes a stack of 3x5 index cards; each card containing one procedure relating to administration or security. After each technician completes the prioritizing of his stack of cards, the group of technicians assembles and rates that one group's set of cards which becomes the final listing of procedures for that section of the design specification. The technicians become familiar with the consensus-seeking process prior to completing this exercise. A short consensus-seeking exercise is provided in order to acquaint them with this process.

The equipment area of the test involves the technicians providing inputs via working in sets of prepared booklets. Each set of booklets represents an equipment subsystem identified as one for which the NEC is responsible. The purpose of this exercise is for the technician to provide insights into the relative importance and difficulty which an entry-level or minimally competent technician should find with that equipment. The group of technicians is presented with an overview of the description of the minimally competent technician, in terms of patrol experience, rating, and submarine/watchstander qualifications.

The first equipment booklet provides for a hierarchialy ranking of the equipment subsystems. It is here that the technician ranks the equipments according to categories of performance of maintenance procedures. Based on a summation of categories of performance of each of these ratings, a hierarchial listing of equipment subsystem is produced. The lowest of equipments might be deleted from the test, depending on the number of subsystems of equipment associated with that NEC. It is important to note that of the approximately 270 items allocated to equipments, a proportional number of items will be assigned hierarchially to each equipment.

Having completed the equipment hierarchial ranking, the technicians next move into the rating of individual equipment packages. This exercise will provide the necessary inputs to determine which phase of equipment

operation should be tested. The technicians will review each operating,standard maintenance, casualty, and corrective maintenance procedure associated with each equipment package. These ratings will be in terms of criticality of performance and difficulty of performance. Immediately after the packages are rated, a computerized scoring algorithm will process the data and produce a hierarchially ranked listing of jobs/tasks associated with each equipment subsystem, and its ranking with reference to the entire subsystem.

This listing will also contain the numbers of items to be assigned to each of the jobs/tasks or procedures. The next step for the group of technicians is to assign suggested cognitive levels for testing purposes to each of the jobs/tasks or procedures. This cognitive level exercise is preceded by a brief workshop in the process of determining appropriate cognitive levels.

Prototype Test Development

The finalized test design specification is produced on-site via a word processing computerized process, and is reviewed by the technician group prior to the end of the workshop. The next step is for the contractor to gain approval of the test design specification by SSPO and the Type Commander. After approval is granted, a prototype test is developed by the contractor. This test, which reflects 360 multiple-choice type questions, follows the design specification. Question items which are specified a cognitive levels above recall are often written in situation sets to elicit responses that would reflect understanding of problem solving associated with the equipments or administrative procedures.

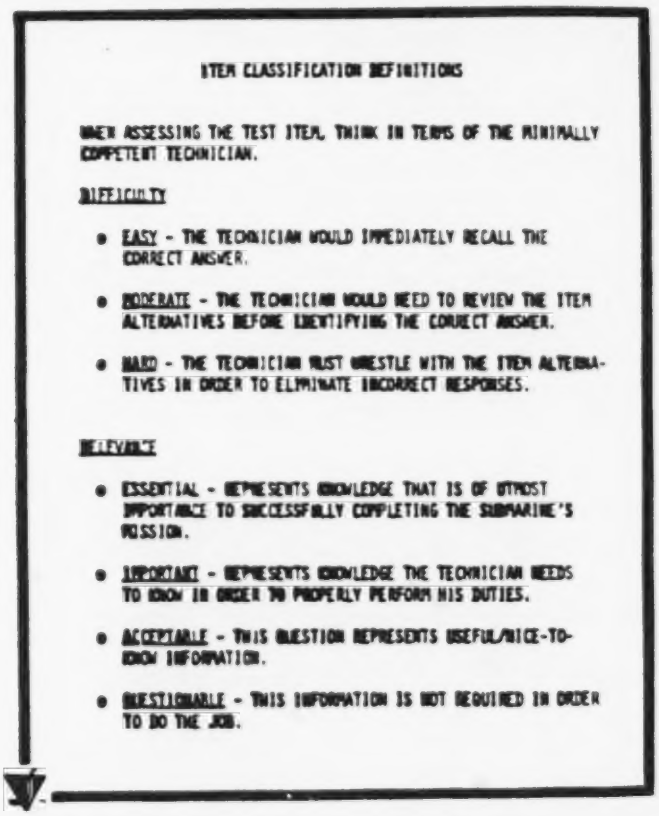
Performance Standard Development

A second week-long workshop is held after the prototype test is developed and after an internal quality control process is accomplished. This Q-C process involves reviewing the items for compliance with a NAVY OD specification developed for the FBM/SWS program. The second workshop involves another two groups of NAVY technicians representative of the NEC; one at the watchstander and one at the supervisor level. This second group of technicians, again working independently but concurrently with each other, will have the task of reviewing their respective test and determining acceptable performance standards. The task for the technician group is to ascertain the minimally acceptable performance level for the watchstander and supervisor respectively. The process adopted and adapted for this ascertaining of minimally acceptable performance standards is a modified Ebel Method. This modified Ebel was developed after consultation with Educational Testing Service consultants and other consultants, all of whom had extensive experiences in the testing and measurement sciences. The Ebel Method involves a jury panel of job incumbents and others with an expertise in the job under consideration, all of whom will make judgments concerning the ability of a minimally competent candidate to answer each question item or series of items correctly.

In practice, an identification of a committee of successful job incumbents and supervisors of incumbents is selected. The Submarine Groups have this as a tasking. Questionnaire instruments are prepared to elicit responses from each of the job incumbents. At the outset of the workshop each committee

member is asked to take the prototype test. The tests are then scored, and the results made known to the committee. This has the effect of providing a benchmark of performance to the committee. After all, they realize that they represent the higher performers in that NEC.

The next phase of the workshop involves each committee-member reviewing each area of the test, section by section, from administration/security through each equipment subsystem, which is prepared in the questionnaire booklets. The technician committee-member is asked to respond to the question item in terms of its difficulty (categories include easy, moderate, hard), and its relevance to the overall mission of the boat (categories include essential, important, acceptable and questionable). Relevance ratings represent the following: (a) essential - represents knowledge that is of the utmost importance to successfully performing a task; (b) important - represents knowledge that a technician needs to know in order to properly perform his duties; (c) acceptable - this question area represents useful/nice-to-know information; and (d) questionable - this information is not required in order to do the job.



ITEM CLASSIFICATION DEFINITIONS

WHEN ASSESSING THE TEST ITEM, THINK IN TERMS OF THE MINIMALLY COMPETENT TECHNICIAN.

DIFFICULTY

- EASY - THE TECHNICIAN WOULD IMMEDIATELY RECALL THE CORRECT ANSWER.
- MODERATE - THE TECHNICIAN WOULD NEED TO REVIEW THE ITEM ALTERNATIVES BEFORE IDENTIFYING THE CORRECT ANSWER.
- HARD - THE TECHNICIAN MUST WRESTLE WITH THE ITEM ALTERNATIVES IN ORDER TO ELIMINATE INCORRECT RESPONSES.

RELEVANCE

- ESSENTIAL - REPRESENTS KNOWLEDGE THAT IS OF UTMOST IMPORTANCE TO SUCCESSFULLY COMPLETING THE SUBMARINE'S MISSION.
- IMPORTANT - REPRESENTS KNOWLEDGE THE TECHNICIAN NEEDS TO KNOW IN ORDER TO PROPERLY PERFORM HIS DUTIES.
- ACCEPTABLE - THIS QUESTION REPRESENTS USEFUL/NICE-TO-KNOW INFORMATION.
- QUESTIONABLE - THIS INFORMATION IS NOT REQUIRED IN ORDER TO DO THE JOB.

For each item, the committee-member will assign one of the rankings from each of these two dimensions. After the committeemembers complete this exercise for a test section, which is included in one questionnaire booklet, the panel moderator, a contractor representative, will convene the entire group to discuss the findings and reach a consensus decision and overall ranking for the items in that test section. This consensus ranking is

accomplished using the matrix format depicted below.

IMPORTANCE	DIFFICULTY: EASY	DIFFICULTY: MEDIUM	DIFFICULTY: HARD
ESSENTIAL	# OF ITEMS 10 PROPORTION .90	# OF ITEMS 15 PROPORTION .85	# OF ITEMS 8 PROPORTION .70
VERY IMPORTANT	# OF ITEMS 17 PROPORTION .80	# OF ITEMS 6 PROPORTION .75	# OF ITEMS 3 PROPORTION .60
NOT VERY IMPORTANT	# OF ITEMS 4 PROPORTION .65	# OF ITEMS 8 PROPORTION .50	# OF ITEMS 9 PROPORTION .50

CUT-OFF SCORE = (10 x .90) + (15 x .85) + (8 x .70) +
(17 x .80) + (6 x .75) + (3 x .60) +
(4 x .65) + (8 x .50) + (9 x .50) = 88.4

MAXIMUM POSSIBLE SCORE = 90.0

Pre-established percentages are assigned to each cell of the matrix to assist in converting the findings to a percentage "cut-score" for test reporting. The moderator begins a discussion of the test items and displays the findings offered by each committee member on the graphic. Outlyers, that is major discrepancies from any established pattern of response are discussed. A consensus finding is the goal of the exercise. This procedure is then repeated for each area of the test. After the entire test is reviewed in this manner, the committee has an opportunity to review the findings for each area of the test.

Test Scoring Format

After the cut-scores are established and the prototype test is approved, the SAT is released for administration. The SAT program is designed for at-sea administration. Therefore, a package of tests, proctor guides and ancillary testing materials is prepared and delivered to the crew prior to deployment. When the ship returns to home port, the answer sheets are optically scanned for scoring purposes. The individual SAT is scored by area/subarea of the test. The number of items correct is presented as a percentage correct of the total number of items included in that area/subarea of the test. This information is then presented graphically in a Test Report which is distributed to the SSPO, TYCOM and Crew Commander.

Test Reporting Format

A typical Score Report will present the overall test contents in outline form. It will present, by personnel tested, the recommendations for training as a result of the test. These recommendations will be made by area/subarea of the test. The report will also flag potential problem areas identified as a result of a composite of the test findings, including the crew in question. This is accomplished by presenting both the mean (x) and the average scores, as well as the scores of the crew in question, for each area/subarea. This information is also graphically displayed. A summary sheet will also call out individual scores for each area/subarea for each technician who takes the test on that crew. A score falling below the cut-point will be bracketed for easy recognition.

09/10/85 SYSTEM ACHIEVEMENT TEST SSBN GOLD CH
Nav/0410/01 Tested 08/17/85
MISSILE TECHNICIAN (MEC 3315) — WATCHSTANDER Page 1

TEST CONTENTS

A. ADMINISTRATION/SECURITY
1. Security
2. Administration
B. WATCH QUALIFICATIONS
C. CASUALTY PROCEDURES
D. EQUIPMENT
1. Missile Gas
2. Launcher Hydraulics
3. Missile Heating & Cooling
D. EQUIPMENT (Continued)
4. Missile Dehumidification
5. Weapons Power
6. Launcher Tube Group
7. Control Group
8. Sm Monitoring Equipment
9. Radiation Monitoring
10. Ejector Group
11. C4 Missile
12. Missile Compensation/Monitoring
13. Guidance Handling Equipment
14. 10 Kw Rectifiers
15. 1.2 Kw Rectifiers

TEST PERFORMANCE SUMMARY

AS A GROUP, TECHNICIANS TESTED MET THE MINIMUM REQUIREMENTS OF COMSUBLANTINST 1500.25C.

The following personnel did not meet the minimum requirements for the test areas and/or subareas listed. Training is recommended. Refer to the RELATED TRAINING listing at the end of this report for available training courses, activities, and materials.

PERSONNEL	AREA	SUBAREA
MT2	ADMINISTRATION/SECURITY	Security C4 Missile Guidance Handling Equipment
MT2		Security
MT2		Missile Dehumidification
MT3	ADMINISTRATION/SECURITY	Security Administration
	WATCH QUALIFICATIONS	Missile Gas Launcher Hydraulics Control Group Radiation Monitoring

Summary

This presentation has described a testing program currently in operation for the FBM/SWS Submarine fleet. It discusses the major iterations conducted to arrive at a performance based System Achievement Test. These steps include a test design workshop which involves incumbent technician personnel reviewing documentation for suitable information for testing; developing and acquiring test items at appropriate cognitive levels for inclusion into the test; developing a prototype test and subjecting this test to review; conducting a cut-score or performance standard determination workshop to ascertain appropriate performance levels; and scoring and reporting the findings in a meaningful manner for those officer personnel who must make personnel judgments about the state of readiness of the FBM program.

A Study of Suggestopedia at the Defense Language Institute Foreign Language Center (DLIFLC)

Brian J. Bush
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

This report documents the results of a study conducted to evaluate the effectiveness of "Suggestopedia", a method proposed to accelerate language learning, as compared to the standard instructional methodology implemented by DLIFLC. An additional effort was made to evaluate a flexibly-scheduled methodology identified by DLIFLC as a target of opportunity.

Methodological definitions

The Suggestopedia methodology is characterized by a variety of techniques emphasizing a relaxed and positive learning atmosphere. The instruction is delivered in situational contexts maximizing the use of the oral communicative skills of proficiency. The flexibly-scheduled treatment is similar to the standard DLIFLC methodology, except that the former uses a pacing of the presentation of material based upon group readiness rather than a fixed schedule for the presentation of materials. Both methodologies use the Progressive Skill Integration (PSI) approach which is a functional approach to language teaching that stresses the integration of the various components of language (i.e., pronunciation, grammar, vocabulary, writing systems, etc.) into communication skills. It is a progressive approach in that students progress through a number of stages beginning with the perception of new concepts and culminating with the acquisition of working communication skills.

Objectives

The objectives of the study were to compare the effectiveness of the Suggestopedia method with the standard DLI method, as well as to provide an evaluation of the flexibly-scheduled pacing procedure.

Method

Subjects

The study included forty junior enlisted Army personnel scheduled to begin the Russian Basic Course (RBC) randomly selected and sorted into two sections for each of the Suggestopedia and standard DLIFLC courses.

One section of ten junior enlisted Army and Navy personnel comprised the flexibly-scheduled group. This group was previously identified and in place prior to its incorporation in this study.

Measurement instruments

Effectiveness was measured and evaluated by academic performance and student attitudes toward the methodologies. Three measures of academic performance were used. One was a set of achievement test scores. A second instrument used was a Proficiency Advancement Test (PAT), a combined measure of both achievement and proficiency. The distinction between achievement and proficiency is that achievement measures performance on course materials, while proficiency measures performance with the target language regardless of the course of instruction. The third measure of academic performance used was a face-to-face oral interview which is considered a measure of conversational proficiency only.

Measures of both proficiency and achievement were used in order to provide a better evaluation of the Suggestopedia methodology as compared to the standard and flexibly-scheduled DLIFLC methodologies.

Student attitudes were measured using a pre and posttest instrument and a weekly attitude survey. The pre and posttest addressed student attitudes toward learning Russian and toward learning foreign language in general. The weekly attitude surveys measured student attitudes about themselves while in class, their opinions about the class, and opinions about their instructor(s).

Administration of measurement instruments

The instruments designed to measure academic performance were administered at different intervals in accordance with completion dates projected by the Suggestopedia and standard DLIFLC methodologies. The standard DLI group had the normal fifteen weeks for completion of the same curriculum that the Suggestopedia group had for completion at the end of ten weeks. The flexibly-scheduled group found that they completed the same curriculum in fourteen weeks.

Attitudinal measures followed the same schedule as identified for the performance measures. For example, the posttest was administered at the end of weeks ten, fourteen, and fifteen for the Suggestopedia, flexibly-scheduled, and control groups respectively.

Results

Demographic variables

Tests of equality between groups were conducted on demographic variables considered to have a possible relation to treatment outcomes. The variables examined were as follows: Military rank, military occupational specialty (MOS), age, years of military service, Defense Language Aptitude Battery (DLAB) scores, educational level, prior language training, General Technical (GT) scores, and gender. A pretest measure of attitudes toward learning Russian and toward learning languages in general was also used to check the equality of the three groups. No significant differences were found among the three groups on any of the demographic variables or on a comparison of the pretest results.

Achievement tests

An analysis of variance was conducted on the two main components of the achievement tests, written and oral. Results from the written component indicated significant differences among the Suggestopedia (M=44.63), standard DLIFLC (M=81.49), and flexibly-scheduled (M=80.87) groups, F (2,46)=48.21, p <.05.

Results from the oral component also indicate significant differences among the Suggestopedia (M=62.90), standard DLIFLC (M=78.23), and flexibly-scheduled (M=79.93) groups, F (2,46)=11.43, p<.05.

Results of T-tests indicated that both the standard DLIFLC and flexibly-scheduled groups scored significantly higher than the Suggestopedia group on the oral and written components. There were no significant differences between the standard DLIFLC and flexibly-scheduled groups.

The results indicate that the greatest differences between the Suggestopedia and the two comparable DLIFLC groups are on the written portion of the achievement tests. This finding is expected, in part, because of the emphasis placed by the Suggestopedia methodology on the oral rather than written features of language learning. The comparability of scores for the standard DLIFLC and flexibly-scheduled group reflect both the similarity between their methodologies as well as their balance with written and oral instruction.

Proficiency Advancement Tests (PATS)

An analysis of variance was conducted on each component of the PAT, (i.e., listening, reading, and speaking). No significant difference was found on the listening component among the Suggestopedia (M=58.46), standard DLIFLC (M=66.23), and flexibly-scheduled (M=63.73) groups, F (2,43)=2.83, p>.05.

Significant differences were found on the reading scores among the Suggestopedia (M=66.71), standard DLIFLC (M=73.47), and flexibly-scheduled (M=75.10) groups, F=(2,43)=6.16, p<.05.

The third component of the PAT, speaking, again reflected significant differences between the Suggestopedia (M=.64), standard DLIFLC (M=.89), and the flexibly-scheduled (M=.79) groups, F(2,43)=7.632, p<.05.

T-tests conducted between the groups on each component indicated that, with one exception, the standard DLIFLC and flexibly-scheduled groups scored significantly higher on each measure than the Suggestopedia group. On the listening component there were no significant differences between the flexibly-scheduled and Suggestopedia groups. In each test there were no significant differences between the standard DLIFLC and flexibly-scheduled groups.

As expected, the PAT results, while still showing significantly lower scores for the Suggestopedia group, indicated somewhat higher overall scores on the PAT then on the achievement tests. The PAT is the first measurement which begins to incorporate elements of proficiency, which are more closely in consonance with Suggestopedia's total emphasis towards language learning.

Face-to-face oral interview scores

The analysis of variance conducted on the face-to-face oral interview scores found no differences between treatments. This is most probably attributable to the small level of proficiency attained after only ten or fifteen weeks of study as well as the early emphasis Suggestopedia places upon proficiency as compared to the more gradual development of proficiency, by DLIFLC, through the integration of the various components of language into communication skills and proficiency.

Attitudinal results

Attitudinal measures failed to discriminate among treatment groups. However, it was interesting to note that the students had positive attitudes toward their particular methodologies. They also felt confident, across groups, in their ability with the target language.

Discussion

Based upon achievement and proficiency measures, the Suggestopedia methodology did not accelerate learning in the context of this study. In fact, Suggestopedia's gains were significantly less than those found for the standard DLIFLC and flexibly-scheduled groups on both the oral and written measures of achievement, and on the three measures from the Proficiency Achievement Test (PAT), listening, reading, and speaking, with the aforementioned exception on the listening results.

There is some indication that Suggestopedia scores improved as measures of proficiency were increased. Suggestopedia scores were poorest as compared to the two other groups on the achievement only tests. On the PAT, which first began to incorporate some proficiency applications, Suggestopedia scores improved with regard to the other two groups. A further indication of this trend by Suggestopedia to do better as more measures of proficiency are included for evaluation is that there were no significant differences among the three groups on the face-to-face oral interview, a measure of proficiency only. Therefore it seems apparent that Suggestopedia scores would be lower on the achievement-oriented measures which are more heavily stressed during the earlier stages of language learning. A second point shoudl be made regarding proficiency results, especially for the face-to-face oral interviews. Proficiency findings may not be able to provide clear discriminations between methods at the early stages of language training.

The combination of findings from this study, along with the research in related Suggestopedia areas indicate a potential use for at least some of the components found in the Suggestopedia method, for example, the use of incidental learning as a teaching technique; the emphases on relaxation and a positive attitude toward the target language. Martin and Schuster (1974) and Lipsitt and Lolordo (1963) reported a curvilinear relationship between learning and stress. Learning increased as stress increased up to an optimum level after which learning fell off as a function of increases in stress (demonstrating the Yerkes-Dodson Law).

If Suggestopedia were to be used as some form of enrichment adjunct to the established DLIFLC language training program, it should probably be used from one to five weeks. This may be the best time interval since positive student attitudes and the instructor energy required to implement a Suggestopedia program seemed to peak and then dimish after approximately five weeks of intensive application. However, specific components characteristic of the Suggestopedia methodology that are found useful in this study could be incorporated throughout the course of instruction.

References

1. Martin, D.J., & Schuster, D.H. (1977). The interaction of trait anxiety and muscle tension in learning. Journal of Suggestive-Accelerative Learning and Teaching, 2, 63-67.

2. Lipsitt, L.P., & Lolordo, V.M. (1963 August). Interactive effect of stress and stimulous generalization on children's oddity learning. Journal of experimental Psychology, 66 (2), 210-214.

Heat, Chemical Protective Clothing and Sustained Cognitive Performance

Bernard J. Fine and John L. Kobrick
U.S Army Research Institute of Environmental Medicine
Natick, Massachusetts

Previous research has shown that the relative impermeability of the current Army clothing system for protection against nuclear, biological and chemical (NBC) agents can result in potentially hazardous or incapacitating conditions within the system, particularly during strenuous physical work in the heat (5,6,9,10). There has been very little research on the psychological performance of soldiers in NBC clothing who are working on jobs not requiring strenuous physical activity, and no research on the effects of heat on them.

Heat has been shown to affect cognitive performance markedly in sedentary soldiers, particularly during sustained operations (2,3). The few studies that have investigated sedentary soldiers performing in NBC clothing (e.g.7,8) are lacking in important scientific and methodological considerations, and have not been concerned with sustained performance.

The purpose of this study was to examine definitively the effects of moderate heat on the sustained cognitive performance of sedentary soldiers wearing NBC protective clothing.

Method

Subjects

Twenty-three male soldier volunteers, ages 19-27 (median=21), were studied. Only persons who could read without glasses were acceptable.

Tasks

The tasks, which included aspects of those performed by members of artillery fire direction centers (FDC'S), forward observers and Army communications personnel, were as follows:

(1) Computation of "Site:" - "Site" is an aiming adjustment used by FDC'S when firing. Tape-recorded data were transmitted to the men over headsets in a format similar to that of artillery fire missions. The men recorded the data on a standard form, performed arithmetic calculations, entered data into and read answers from an artillery slide rule and recorded the answers.

(2) Receiving and decoding map grid coordinates: - Pre-recorded, coded coordinates were transmitted as radio messages via headsets. The men recorded the alpha-coded coordinates on a form, chose the correct one of three decoding wheels, translated the code into numeric format and recorded the answers.

(3) Receiving and decoding messages: - Pre-recorded, coded messages of from five to eight words were transmitted as radio messages over headsets. The men recorded each message on an appropriate form, decoded it by referring to a simulated Army codebook and recorded the transcription on the form.

(4) Plotting targets and determining range and deflection: - Each man was given a map with battery positions and deflection reference points. He also had a book with lists of targets (grid coordinates) to plot. He plotted them, found their ranges and deflections from designated batteries and recorded the answers. He also included the time at which he had completed processing each target, thus enabling assessment of number of plots per unit of time.

BEST COPY AVAILABLE

The first three tasks were paced by the rate/frequency of the messages and were not under the mens' control. The map task was at times paced by the study requirements and at other times was "self-paced." The men did not know which of the three kinds of radio tasks they would be required to perform until a message arrived. Messages were designed to mimic real military radio transmissions, including a variety of voices and transient background noises.

In audition to the above tasks, a visual field surveillance task and an auditory perception task were performed during hours 2,4 and 6. The results of these tasks will be reported elsewhere. A number of personality measures and vision tests also were administered, but results are not reported here.

Design and Procedure

The men arrived at three-week intervals in six-man groups. Each group completed its assignment before the next arrived. The groups underwent two weeks of intensive training followed by an "experimental" week to evaluate performance in the heat while wearing the NBC clothing. The clothing system worn was MOPP IV (Mission Oriented Protective Posture; "IV" refers to total encapsulation - suit, worn completely closed, boots, gloves, mask and hood.)

The men trained in a classroom, six to seven hours a day, for two weeks; no week-ends. Training in the codewheel, codebook and Site tasks began with a simple written format, and became progressively more complex until the men could handle the rapid, noisy "military" messages. During training, the men practiced several hundred messages with immediate feedback and discussion of errors. Emphasis was placed first on accuracy, then on speed of performance. Map plotting was practiced for hundreds of trials with immediate feedback of errors. Again, emphasis initially was placed on accuracy rather than speed.

During the first week, the men were briefed on proper procedures for wearing the NBC clothing and performed briefly while wearing single components of the system, e.g., gloves only or mask only. During the second week, they performed the tasks daily with and without MOPP IV. By the end of the week, each had worked in MOPP IV for about eight hours, spread over five days.

The "experimental week" proceeded as follows:

Monday:- Two one-hour "refresher" runs to bring the men up to pre-weekend performance levels on the various tasks (21.1 degrees C.,35%rh);

Tuesday:- Control Day, seven hours at 21.1 degrees C.,35%rh, battle dress uniform (BDU), referred to as "BDU-Control-1;"

Wednesday:- Mopp Control Day, seven hours at 12.8 degrees C.,35%rh, MOPP IV worn over BDU, referred to as "MOPP-Control;"

Thursday:- Control Day, same as Tuesday, "BDU-Control-2;"

Friday:- Heat Stress Day, seven hours at 35.0 degrees C.,40%rh, MOPP IV worn over BDU, referred to as "MOPP-Heat-Stress."

The 12.8 degree C. temperature of the MOPP Control Day was calculated by Breckenridge (1) as the MOPP IV equivalent to the 21.1 degree C. condition used for the BDU for a seven-hour period. These conditions were also used where appropriate during the two-week training period.

The tasks were presented as one-hour blocks of messages, 25 per hour. Five were irrelevant (had addresses to which the men had been trained not to respond), six were codebook, six codewheel and eight Site. Intervals between

932 BEST COPY AVAILABLE

messages ranged from 30 seconds to over two minutes according to a random pattern. There were no duplicate messages throughout the entire experiment.

Messages were sent in random order. Each of the six men had a different order, which was the same from hour to hour, e.g., if message #5 was codebook for man #1 in the first hour, it was codebook for him in all hours. Since messages and intervening intervals varied in length, work patterns differed; the men were forced to work independently, yet on identical material.

The radio messages were presented to the men four times on each of the four experimental days, as hours 1,3,5 and 7.

Concurrent with working the radio messages, the men did the map task. The radio messages had priority; map work was interrupted upon hearing a message, to be resumed only when what was required by the message was completed. Thus, for each of hours 1,3,5 and 7, the men were always engaged in cognitive work.

During each of hours 2,4, and 6, the men worked only on the map task, at their own pace and without interruption, for about 35 minutes. They also were tested for auditory perception, in groups of three, and, individually, for peripheral vision. A ten-minute break was given each hour.

On the NOPP-Heat-Stress Day, rectal temperatures were taken every five minutes or less. Regulations required removal from heat if temperatures reached 39.5 degrees C. The men were encouraged to drink water ad lib; masks and canteens were equipped for drinking in NOPP IV. No lunch was eaten during the experimental week. Access to a portable toilet was permitted only on an emergency basis. Only three men availed themselves of this opportunity during the entire study.

Results and Discussion

Twenty-three men actually arrived to participate. Of those, three were disqualified; one was allergic to the NOPP clothing, one had insufficient training due to illness and one went on sick-call on the final (stress) day. Therefore, all data analyses are based on n=20.

Responses to the messages were scored/verified by criteria established beforehand. Errors were classified into "omission" and "commission." Omission consisted of missing part of a message or translation. Commission included incorrect reception of messages and incorrect translations and computations.

On the NOPP-Heat-Stress Day, two men had to be evacuated for medical reasons. They were given the maximum number of omission errors possible for the messages and scored as having plotted no targets for the time they missed. The group averages reflect those decisions [see Fine & Kobrick (2 page 121)].

CODEBOOK: Results for the codebook task are shown in Figure 1 for errors of omission and commission combined. ANOVA yielded a significant main effect for Conditions ($F=11.75$;d.f.^{3,},^{304};$p<.00001$). No significant differences between the two NBC-Control conditions or between hours of testing within either of the control conditions were evident at any time.

The NBC suit by itself (NOPP-Control) seemed to cause a decrement in performance. Not significant statistically for the first three hours, the effect

BEST COPY AVAILABLE

became stronger after five hours. However, after seven hours, two distinct patterns emerged: performance in the MOPP-Control condition improved to the level of the BDU-Control conditions, whereas the group showed a statistically significant decrement in excess of 30 percent when in the MOPP-Heat-Stress condition. The decrements in performance in both MOPP conditions were found to be due almost entirely to increases in errors of omission.

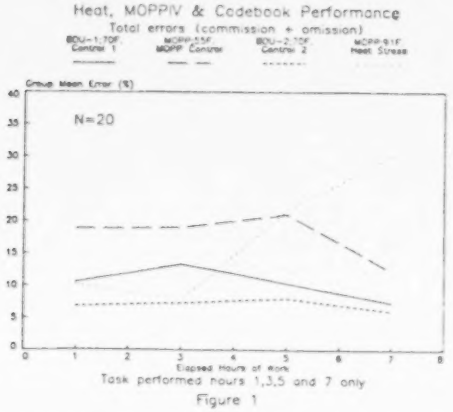


Figure 1

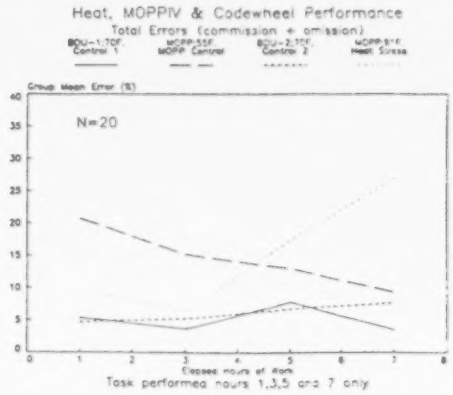


Figure 2

CODEWHEEL: Results for the codewheel task are shown in Figure 2. ANOVA yielded a significant main effect for Conditions ($F=14.28$;d.f.3,304;$p<.00001$), and a significant Conditions by Hours of Work interaction ($F=4.11$;d.f.9,304;$p<.001$). Performance in the BDU-Control conditions showed remarkable consistency and stability. The opposite trends of the two MOPP conditions clearly were evident again. In the MOPP- Control condition, performance initially was adversely affected (significantly only after the first and third hours), only to gradually improve until, after seven hours, it approached the levels of the two BDU conditions. The MOPP-Heat-Stress condition showed no initial adverse effects, but dramatic and significant increases in error occurred from the third to the fifth and the fifth to the seventh hours. Virtually all of the increase occurred as errors of omission.

COMPUTATION OF SITE: This was the only radio task in which the message was repeated; in both other tasks, the message was heard only once, hence the lower error rates for this task, as shown in Figure 3. ANOVA showed a significant main effect for Conditions ($F=7.26$;d.f.3,304;$p<.0001$) and for Elapsed Hours of Work ($F=2.82$;d.f.3,304;$p<.05$) and a signficant Conditions by Hours interaction ($F=3.00$;d.f.9,304;$p<.002$).

Performance in the two BDU-Control conditions again was highly stable, remarkably similar and quite free of errors. The group in these conditions averaged only 2-3% errors consistently for each of the four hours in which the task was performed. There was no significant decrement in performance in the MOPP-Control condition at any time. The MOPP-Heat-Stress condition showed significant increases in percent group error from the third to the fifth and the fifth to the seventh hours, culminating in about a 20% decrement, due entirely to an increase in errors of omission.

NUMBER OF TARGETS PLOTTED, HOURS 1,3,5 AND 7: ANOVA yielded a significant main effect for Conditions ($F=11.4$;d.f.3,304;$p<.00001$), and a near significant effect for Elapsed Hours of Work ($F=2.23$;d.f.3,304;$p<.10$). More targets tended

934 BEST COPY AVAILABLE

to be plotted in the BDU-Control-2 condition than in BDU-Control-1, but this was significant only for the third hour. More important was the tendency for target plotting productivity to be lower in the MOPP conditions, but this was not consistently statistically significant; it did not hold when both BDU control groups were compared with both MOPP groups. The effect was much more apparent when the data for self-paced hours 2,4, and 6 were considered.

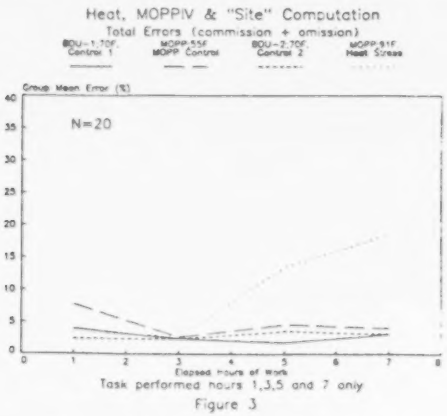
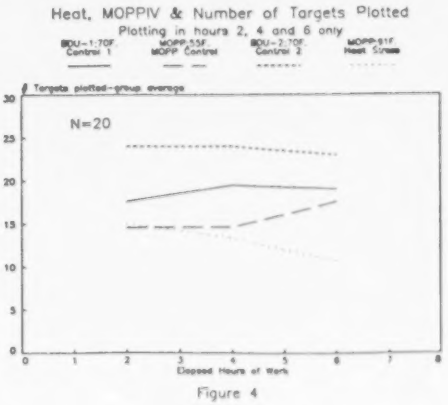


Figure 3


Heat, MOPPIV & Number of Targets Plotted
Plotting in hours 2, 4 and 6 only
BDU-1.70F Control 1
MOPP-55F MOPP Control
BDU-2.70F Control 2
MOPP-91F Heat Stress
# Targets plotted-group average
N=20
Elapsed Hours of Work
Figure 4

NUMBER OF TARGETS PLOTTED, HOURS 2,4, AND 6: These data are shown graphically in Figure 4. The ANOVA resulted in a very significant Conditions effect ($F=31.4$; d.f. 3,228; $p<.00001$). As in hour 3, the men were more productive in the BDU-Control-2 condition than in the other conditions. Here, however, the effect was significant for each of hours 2,4, and 6.

Concerning the effects of the NBC suit, by the fourth hour, the group, when in either MOPP condition, had significantly lower productivity than when in either BDU-Control condition. By the end of the sixth hour, however, productivity in the MOPP-Control condition had returned to the same level as the BDU-Control-1 condition, whereas productivity in the MOPP-Heat-Stress condition showed a significant deterioration (10.7 targets per hour, compared with 17.6 per hour for MOPP-Control, 19 per hour for BDU-Control-1 and 23 per hour for BDU-Control-2). The stability and magnitude of performance in the BDU-Control conditions should be noted, attesting both to the effective training procedures and the high level of motivation of the men.

OTHER PERFORMANCE MEASURES: Several other measures were obtained but were not used: (a) Number of plotting errors- After training, all men could plot with virtually no errors. This superior performance showed in the experimental data. During hours 1,3,5 and 7, the group averaged .33 errors per hour. During hours 2,4 and 6, it averaged .51 errors per hour. Because of the low incidence of errors, statistical analyses were not performed; (b) Number of ranges calculated- The number of ranges was the same as the number of targets since a range was required for each target. Thus, results for ranges and for targets are the same; (c) Range calculation errors- These were so infrequent that statistical analyses were meaningless; (d) deflection calculation errors- Due to an error in placement of deflection reference points, deflections for some targets could not be determined. Since those targets did not occur in a systematic manner, and since each man worked at his own pace, meaningful hourly productivity or accuracy information could not be determined.

935

BEST COPY AVAILABLE

CONCLUSIONS: After four to five hours of exposure to a moderately hot environment, the cognitive performance of a group of highly trained soldiers clad in the MOPP IV configuration of NBC protective clothing began to deteriorate markedly. By the end of seven hours of exposure, increases in percent group error ranged from 17% to 23% over control conditions on investigator-paced tasks. Virtually all of this decrement was due to increases in errors of omission. The productivity of the group on a self-paced task (map plotting) diminished by about 40% from control conditions after six hours in the heat, but accuracy of plotting did not appear to be markedly affected.

While generalization from a small sample should be made with caution, it should be noted that the control data in this study are very similar to control data from three previous studies using the same tasks (2,3,4). This consistency of baseline performance with n=90 lends credence to the validity of the results presented here.

Finally, this study was done in a climatic chamber under conditions in which the men knew that they were being monitored for health and safety. The additional stress of being exposed to actual toxic agent or even the threat of it and/or other hostile situations in all likelihood will have further negative impact on group performance.

References

1.Breckenridge, J.R. U.S. Army Research Institute of Environmental Medicine, Natick, MA. Personal communication, 4 Nov 83.

2.Fine, B.J. & Kobrick, J.L. Effects of altitude and heat on complex cognitive tasks. Human Factors, 1978, 20, 115-122.

3.Fine, B.J. & Kobrick, J.L. Effects of altitude and heat on complex cognitive tasks,II. Unpublished research.

4.Fine, B.J. & Kobrick, J.L. Effects of simulated rapid translocation and heat on sustained performance of cognitive tasks. Unpublished research.

5.Garrett, Jr., P. & DeBellis, W. Human Engineering Laboratory Aviation Supply III/V (HEVAS III/V). Presented at Tri-Service Aeromedical Rsch. Panel Chemical Defense Meeting, San Antonio, TX, 1981.

6.Goldman, R.F. & Breckenridge, J.R. Current approaches to resolving the physiological heat stress problems imposed by chemical protective clothing systems. In Proc. 1976 Army Sci Conf., West Point, N.Y. Vol. IV,447-453.

7.Hamilton, B.E., Simmons, R.R. & Kimball, K.A. Psychological effects of chemical defense ensemble imposed heat stress on Army aviators. U.S. Army Aeromedical Rsch. Lab., Ft. Rucker, AL, USAARL Rept. #82-7, 1982.

8.Hamilton, B.E. & Zapata, L. Psychological measurements during the wear of the U.S. Aircrew Chemical Defense Ensemble. U.S. Army Aeromedical Research Lab., Ft. Rucker, AL, USAARL Rept. #83-7, 1983.

9.Martin, H. DeV. & Goldman, R.F. Comparison of physical, biological and physiological methods of evaluating the thermal stress associated with wearing protective clothing. Ergonomics, 1972, 15, 337-342.

10.Myrhe, L.G. Selection and physical conditioning requirement of rapid runway repair personnel with the CB ensemble. Human Engr. Lab. Aviation Supply III/V (HEVAS III/V). Presented at Tri-Service Aeromed. Rsch. Panel Chemical Defense Technical Meeting, San Antonio, TX, 1981.

(Authors adhered to AR 70-25 and USAMRDC Reg. 70-25 on Use of Volunteers in Research. Views, opinions and/or findings are the authors' and do not reflect official Dept. of Army position, policy or decision.)

BEST COPY AVAILABLE

Exceptional Recruits: A Look at High- and Low-Aptitude Personnel

Anita R. Lancaster
Office of the Assistant Secretary of Defense
(Force Management and Personnel)

The Department of Defense, on a quarterly and annual basis, reports on the "health of recruiting" by describing the quality and quantity of youth accessed. Recruit quality is defined in terms of aptitude test scores and educational level (high school graduation status). During periods of favorable recruiting, such as that experienced since 1981, DoD typically reports that accessions have above average aptitude and that their high school graduation status is as good or better than that of the American youth population.

Recently, however, DoD has reported recruiting men and women of unprecedented quality. For example, in Fiscal Year (FY) 1984, all four Services met or exceeded their accession objectives. Of the 333,400 young men and women entering the Services last year, 93 percent possessed high school diplomas and 93 percent scored average or above average on the enlistment aptitude test. This contrasts with previous high school graduate recruit levels of 91 percent in FY 1983, 86 percent in FY 1982, 81 percent in FY 1981, and 68 percent in FY 1980. Similarly, the FY 1984 aptitude level of 93 percent in the average or above category compares with 92 percent for FY 1983, 87 percent for FY 1982, 82 percent for FY 1981, and 69 percent for FY 1980. Although the trend has been upward, it appears that recruiting results for FY 1985 will be similar to those of FY 1984.

Seventy-five percent of American youth have high school diplomas and 69 percent of a nationally representative sample of American youth ages 18 to 23 scored average or above on the enlistment test. Given that 93 percent of FY 1984 recruits both have high school diplomas and scored average or above on the enlistment test, it would seem that the Services have taken in a disproportionate share of high quality American youth into their ranks over the past five years.

Because of the exceedingly high quality levels achieved in recent years, there has been considerable speculation that the Services' abilities to successfully enlist such youth, over time, simply cannot continue. In fact, some contend that an improving economy, a decreasing manpower pool, and keen competition with higher education and employers for bright young people may result in a significant decline in the quality of recruits in the future.

There is no doubt that if recruiting becomes more difficult, the costs of competing with the private sector for exceptionally high quality youth must be weighed against the consequences of incrementally lowering recruit quality levels. Someday, for example, we may have to consider the option

of lower aptitude standards against increased military pay or reenlistment bonuses. There also, undoubtedly, is some point of diminishing returns - where a decline in recruit quality would adversely affect force readiness. Current state-of-the-art, however, does not provide us with any precise way to calculate the effects of varying such options.

It is incumbent upon the Department of Defense to begin to plan for potential recruiting difficulties now. It would be unconscionable for us to believe that current recruit quality will continue in an unmitigated flow. Similarly, we would be remiss in not exploring the development of methodologies for determining appropriate quality levels so that we do not increase recruiting resources to maintain inflated levels. We would not want to make these decisions capriciously. We need to know how much it costs to recruit, train, supervise and retain personnel at all aptitude levels in order to make informed decisions.

Fortuitously, we are in a position to study the effects of having accessed both high- and low-aptitude recruits. At last year's Military Testing Association conference, Lieutenant General E. A. Chavarrie, Deputy Assistant Secretary of Defense (Military Manpower and Personnel Policy) outlined the research efforts we then were initiating in this area, and I am pleased to report that we have made considerable progress since then.

With regard to researching the effects of enlisting high-aptitude personnel, Ms. Janice H. Laurence, from the Human Resources Research Organization (HumRRO), will present a paper she co-authored with Ms. Elizabeth Schneider, also from HumRRO, on the "Enlistment and Utilization of High-Aptitude Recruits." This competitively bid project recently was awarded to HumRRO and, through this research, we will be learning about the demographic characteristics and pre-service experiences of high-aptitude recruits. In addition, HumRRO researchers will be examining the occupational assignment of high-aptitude personnel and their performance, relative to lower aptitude personnel; thus, we hope to be able to develop a model for estimating the incremental benefits of enlisting higher aptitude youth versus those of somewhat lower aptitude. If we able to do so effectively, we will be in a position in the future to answer "What...if" types of questions with regard to the effects of lowering or raising recruit standards.

Another HumRRO researcher, Dr. Barbara M. Means, will be presenting a paper she co-authored with Ms. Arti Nigam and Ms. Jane G. Heisey, also from HumRRO, entitled, "When Low-Aptitude Recruits Succeed." These researchers have been studying the characteristics and performance of two cohort groups - those recruits accessed during Project 100,000, when low-aptitude personnel deliberately were enlisted, and the recruits who were unintentionally accessed during the misnorming of the enlistment test in the late 1970s. These analyses involve not only an examination of the military performance of low-aptitude recruits, but an assessment of the impact of military service on the post-military lives of these recruits. The data that Dr. Means will share on the military performance of low-ability personnel is illuminating.

Finally, Dr. W. S. Sellman, Director for Accession Policy, Office of the Assistant Secretary of Defense (Force Management and Personnel), will provide us with discussant remarks. Dr. Sellman was one of the original Project 100,000 researchers under Mr. I. M. Greenberg, then Director of Project 100,000 for Defense Secretary Robert S. McNamara and later a Deputy Assistant Secretary of Defense. Today, Dr. Sellman has responsibility for setting policy for recruitment and utilization of high- and low-aptitude personnel.

Using Military High- and Low-Aptitude Recruits Wisely

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Force Management and Personnel)

Time and time again the Department of Defense has expressed its interest in and need for quality personnel. We try to maximize the enlistment of youth whose aptitude and education level ensure the highest probability of successful military performance. Nongraduates and lower-aptitude personnel are less adaptable to military life and are harder to train than their high school graduate, higher-aptitude counterparts.

Despite DoD's preference to enlist the bright and well educated, recruiting market trends, conditions, and constraints in the past have led to enlistment of substantial numbers of low-aptitude personnel and less than optimal use of high-aptitude personnel. Though currently there is no recruiting "red alert" in terms of attracting adequate numbers of bright young people into the military, prudent management calls for planning against a time when this might not be true.

With a declining manpower pool, the Services will be in heavier competition with the private sector for young "employees," particularly those of high-aptitude. Because high-quality personnel may become more difficult to recruit and retain, it is a propitious time to begin studying in more depth both the performance of low- and high-aptitude personnel so that we may use our resources wisely.

Lessons from the Past

As mentioned previously, the Defense Department has experience with accessing large numbers of relatively low-aptitude personnel. From 1966 through 1971, for instance, 320,000 low-aptitude men entered service under "Project 100,000." This program, as part of President Lyndon Johnson's War on Poverty, admitted low-aptitude men, including many who would have been disqualified under previous aptitude (test score) standards, into the military so that they might learn useful skills. Though Project 100,000 was initiated to provide the military's educational and training opportunities to our nation's culturally disadvantaged youth, it also served as a partial solution to the quantitative manpower demands imposed by increasing American involvement in Vietnam.

Between 1976 and 1980, a second large-scale accession of low-aptitude individuals occurred. This occasion was not in response to a specific, planned program; rather these large numbers of low-aptitude recruits resulted from the inadvertent misnorming of the Armed Services Vocational Aptitude Battery (ASVAB), the test given to select and classify recruits (Sellman & Valentine, 1981). Had the test been properly calibrated, many of these "Potentially Ineligibles," as Greenberg (1980) named them, would not have qualified for enlistment or would have been assigned to less demanding

occupational specialties. This "unfortunate" occurrence may have also been fortuitous in forestalling major manpower shortages during this difficult recruiting period.

Looking Toward the Future

Many people believe that the Defense Department's role as an institution in American society goes beyond that of "providing for the common defense." Because the military is this country's largest employer of youth, many social engineers believe that DoD should have responsibilities for training and imparting lasting skills and values for society's disadvantaged youth. Military Service, they believe, should provide job-related training opportunities for valuable social and growth experiences, and increased interaction and cooperation with a wide variety of different kinds of people. Disadvantaged youth should be afforded the opportunity to "grow-up" in a relatively structured environment--learning to be responsible productive citizens.

One does not have to agree with this role for DoD to understand that the Services have accessed low-aptitude youth many times to come to the nation's defense and to provide needed manpower. In the event that the Defense Department is called upon in the future to admit large numbers of low-aptitude recruits, the lessons learned from the past may prove invaluable. DoD should understand more fully the effects of low-aptitude personnel on the military and the military's effect on such individuals.

From Low to High: Developing Human Resources

The Department of Defense's role as a developer of human capital has been directed primarily at those recruits who are of average or above average ability. Less is known about developing the skills and abilities of low and very high-aptitude recruits so that they might reach their respective full potential in the military and carry with them the benefits of their military experience into civilian life.

Three research studies are currently being conducted which address these questions. Our project dealing with high-aptitude enlistees will describe their enlistment motives, performance across jobs of varying complexity, and reasons for leaving military service. This study will also compare the costs and benefits associated with accessing high- versus average-aptitude enlistees. This is the most comprehensive attempt to investigate performance of high-aptitude personnel.

The second study aims at investigating the ability of low-aptitude individuals to function effectively in the military--an issue of concern for years. Though research indicates that low-aptitude enlistees do less well in military training, there is a scarcity of empirical data concerning their actual job performance. Regardless of DoD's role as a social institution, information is needed concerning the costs of training and supervising these individuals when placed in various military occupational categories.

The third study examines the effect of military service on the post-military lives of low-aptitude veterans. Advantages provided to low-aptitude enlistees (i.e., maturation, identification with an established organization, enhanced self worth, responsibility, training opportunities) are obvious, according to social observers. However, actual data documenting employment and social experiences of low-aptitude veterans has not been available generally. This project allows for detailed comparisons of low-aptitude veterans accessed during Project 100,000 (1966-1971) and the ASVAB misnorming (1976-1980) with their low aptitude nonveteran counterparts.

DoD's current pioneering efforts, aimed at studying the utilization and performance of youth at both ends of the ability distribution, will enable us to use our precious resources wisely. Knowing the strengths and weaknesses of both high- and low-aptitude personnel will enable DoD to assign them more efficiently and better meet the challenges of the changing demographics of the next decade.

References

Sellman, W.S., & Valentine, L.D. (1981, August). Aptitude testing, enlistment standards, and recruit quality. Paper presented at the 89th Annual Convention of the American Psychological Association, Los Angeles, CA.

Greenberg, I.M. (1980). Report on mental standards for enlistment: Performance of Army personnel related to AFQT/ASVAB scores. In Department of Defense, Implementation of new Armed Services Vocational Aptitude Battery and actions to improve the enlistment standards process. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, & Logistics).

When Low-Aptitude Recruits Succeed

Barbara M. Means, Arti Nigam, and Jane G. Heisey

Human Resources Research Organization

> The CNO...does not want enlistment standards lowered and he certainly does not want another "Project 100,000," the infamous effort of a generation ago to bring into the armed services individuals below minimum enlistment standards and "train them up" so they could perform adequately. "Project 100,000" is almost universally remembered by military people as a disaster that caused the military services tremendous grief.
>
> John Burlage, The Navy Times, February 25, 1985

> ...there was little, if any, measurable adverse effect on the services and very substantial gains, in my opinion, to the individuals and to society.
>
> Robert S. McNamara, in interview with T. Sticht, April 26, 1985.

The extent to which the military services should accept individuals with low aptitude test scores for military enlistment has been, and remains, a controversial topic. Both during Project 100,000, when the acceptance of large numbers of young men previously ineligible because of their aptitude scores was instituted as a social welfare program, and during the misnorming of the enlistment examination when thousands of low-aptitude individuals were accidentally accepted for military service, emotions have run high. Military commanders argue that increased training costs, poorer military performance,

---

Paper presented at the Military Testing Association, San Diego, October 1985.

This research was performed under contract No. N66001 with the Navy Personnel Research and Development Center and was sponsored by the Directorate for Accession Policy, Office of the Assistant Secretary of Defense (Force Management and Personnel). The views, opinions, and findings expressed in this paper are those of the authors and should not be construed as Department of the Navy or Department of Defense position or policy unless so designated by other documentation.

The authors wish to extend their gratitude to the Defense Manpower Data Center (DMDC) and particularly to Les Willis, who performed the statistical tabulations of DMDC Cohort File data.

and a tarnishing of the military's image result from accepting large numbers of such individuals. Social critics point to the military's capability to provide a positive force for improvement in the lives of the economically and educationally disadvantaged, and note that when operational aptitude standards are raised in good recruiting years, smaller proportions of women and minorities (who score lower on the enlistment examination) are accepted for military service.

For enlistment purposes, intellectual aptitude is measured by Armed Forces Qualification Test (AFQT) scores, which are reported in terms of percentiles and are grouped into five broad categories--Categories I through V--with Categories III and IV further divided into subgroups denoted by letters. Numerically higher groups signify lower aptitude. Category IVA, for example, includes percentiles 21 through 30, and was the lowest eligible aptitude category for high school graduates in the Navy, Air Force, and Marine Corps; the Army accepted high school graduates in Category IVB, which includes percentiles 16 through 20. Higher aptitude standards were set for non-high school graduates (except in the Marine Corps).

Rather than join either side of the "I loved it" - "I hated it" debate, we have chosen to reframe the question. When low-aptitude individuals are brought into military service, where do they perform best? The services, of course, have their own ideas about this question, and these are reflected in the minimum aptitude scores set for entry into various jobs and in quotas set for the maximum number of Category IV personnel that can be recruited for a given job. A July 1984 memo from the Army Deputy Chief of Staff for Personnel (DCSPER) to the Army Recruiting Command, for example, stipulated that the specialties for which the most Category IV personnel could be recruited that year were infantry (14.4% of Category IV recruits), motor transport (8.4%), cannon crew (6.1%), and food service (5.1%). Many jobs, such as intelligence analyst, radar controller, and computer repairer were not to get any Category IV personnel.

The largest, recent data set concerning the performance of low-aptitude military personnel is found in the Defense Manpower Data Center (DMDC) records for those individuals who entered service between January 1976 and October 1980 as a result of an enlistment test miscalibration. The particular aptitude score range of those individuals let in by mistake varies depending on the service entered and the individual's education level, since required scores vary for these groups. The analyses to be reported in this paper contrast the highest of the ineligible aptitude categories (which contained most of the inadvertent acceptees) for each service/education group with the lowest of the aptitude-eligible groups. Because military jobs differ considerably in terms of their duties and required intellectual aptitude, all performance measures were examined separately for the nine DoD occupational codes.

Although there is considerable variation within each occupational code in terms of the aptitude composite scores stipulated for entry into various jobs within that code, the occupational codes do vary considerably in terms of average score required and thus can be classified according to aptitude demands as perceived by military managers. The codes with high requirements are electronic equipment repair (1), communications and intelligence (2),

medical and dental technicians (3), and other technical and allied (4). Medium requirements are set for functional support and administrative (5) electrical/mechanical equipment (6), and craftsmen (7) specialties. Low requirements are set for infantry, gun crew, and seamen (0) and for service and supply handlers (8). These job demand categories were used in our analyses of military performance.

Probably the most basic--and the most often quoted--military performance index is attrition. To look at our comparison groups, which contain men who entered in different years and enlisted for different terms of service, we counted as attrition cases all those who had separated from enlisted service for reasons other than entry into an officer program or completion of their term of service. These data are shown for the aptitude comparison groups, by education and occupation type, in Table 1. The data clearly show that if completion of one's term of enlistment is the criterion for success, the inadvertent enlistees did as well as those in the higher-aptitude comparison groups. In fact, the lower-aptitude groups generally had somewhat lower attrition rates, at least in the Army and Navy. Their rates were significantly lower ($p<.01$) in the Army among nongraduates in low- or medium-requirement jobs. Similarly, in the Navy, the aptitude-ineligible group had less attrition in both low- and medium-requirement ratings, regardless of education level. (It should be noted, however, that in the Navy many individuals do not have an assigned occupation, and lower-aptitude individuals were more likely to be in this no-occupation group.) The usual 2:1 difference between nongraduates and graduates in terms of attrition rates is apparent in this table, and the education differences completely overwhelm those attributable to aptitude category. These findings are compatible with those of Shields and Grafton (1983), who found no difference between these aptitude groups in terms of first-term completion for Army enlistees in 18 specialties.

Shields and Grafton examined other first-term performance measures as well. The only measure to show any differences between these aptitude groups--and the differences were modest--was Skill Qualification Test (SQT) score. The Army's SQTs are job-specific tests containing hands-on, ratings, and written test measures of task performance. Most of the tests' variance derives from the written portion of the exam, and hence a correlation with AFQT score is to be expected. Nevertheless, differences were not impressive: averaged over 17 jobs, the mean standardized SQT score was 96 for Category IVA-B (eligible) and 91 for Category IVC (ineligible) accessions among high school graduates; among nongraduates, it was 99 for Category IIIB (eligible) and 95 for Category IVC (ineligible). (The test is standardized with $\bar{X}=100$ and $SD=20$.) In an earlier report, Greenberg (1980) reported a similar pattern of SQT scores among FY 1977 Army accessions across nine jobs.

With the passage of time, we can now look at indications of career performance for many of the individuals who entered during the ASVAB misnorming. Since most of these measures, such as retention and promotion, are dependent upon date of entry, we decided to look at a single cohort--FY 1977 entrants. (This cohort includes an extra quarter because of the change in government fiscal year, and hence is the largest single-year cohort from that period.) Using records that were updated as of December 1984, we could examine military records for over seven years of military service. To permit an unbiased examination of promotion rates, those who entered with an advanced pay grade (anything other than E-1) were deleted from the data set.

Table 1

Percent Attrition for FY 1976 - FY 1980 Male Accessions in Aptitude Comparison Groups, by Occupational Requirement Level, Service, and Education
(N appears in parentheses)

Occupational Requirement	ARMY				NAVY				AIR FORCE				MARINE CORPS			
	HSG		NHS		HSG		NHS		HSG		NHS		HSG		NHS	
	IVC	IVB	IVA	IIIB	IVB	IVA	IVA	IIIB	IVB	IVA	IIIA	II	IVB	IVA	IVB	IVA
Low	19* (18,172)	20 (20,444)	41** (27,595)	43 (25,526)	5** (2,235)	8 (3,357)	13** (1,001)	20 (2,865)	24 (408)	22 (3,015)	41 (1,304)	41 (1,896)	19 (6,819)	19 (10,009)	30 (944)	31 (3,649)
Medium	14 (12,581)	14 (14,703)	30** (15,275)	33 (16,901)	4** (7,287)	6 (13,755)	12** (3,019)	15 (8,562)	20 (1,385)	18 (8,794)	40 (3,678)	38 (5,513)	13 (2,217)	13 (4,479)	18 (256)	21 (1,065)
High	11 (4,312)	12 (7,287)	30* (7,870)	32 (10,470)	7 (1,097)	8 (2,990)	15 (427)	17 (2,102)	16 (141)	15 (1,003)	31 (1,006)	28 (2,941)	14 (656)	13 (1,269)	22 (83)	26 (427)

*Significantly different from higher-aptitude comparison group at $p<.05$.
**Significantly different from higher-aptitude comparison group at $p<.01$.

858

Table 2

Percent Retention as of December 1984 for FY 1977 Male Accessions in Aptitude Comparison Groups, by Occupational Requirement Level, Service, and Education
(N appears in parentheses)

Occupational Requirement	ARMY				NAVY				AIR FORCE				MARINE CORPS			
	HSG		NHS		HSG		NHS		HSG		NHS		HSG		NHS	
	IVC	IVB	IVA	IIIB	IVB	IVA	IVA	IIIB	IVB	IVA	IIIA	II	IVB	IVA	IVB	IVA
Low	18** (4,236)	21 (4,554)	9 (7,030)	9 (5,534)	36* (658)	30 (1,013)	25 (483)	24 (868)	20 (64)	25 (512)	15 (157)	20 (310)	11 (1,612)	11 (2,304)	3 (319)	5 (1,164)
Medium	26 (3,403)	26 (3,867)	15 (5,191)	14 (4,817)	34** (1,737)	30 (3,634)	26 (1,110)	23 (2,181)	35 (170)	35 (1,668)	24 (475)	24 (971)	30 (405)	29 (754)	16 (86)	17 (327)
High	34 (1,028)	33 (1,784)	19 (2,157)	18 (2,474)	42 (243)	37 (673)	37* (145)	28 (497)	--[a] (29)	38 (209)	30 (141)	30 (456)	33 (114)	27 (214)	-- (23)	10 (133)

*Significantly different from higher-aptitude comparison group at $p<.05$.
**Significantly different from higher-aptitude comparison group at $p<.01$.
[a]Fewer than 50 cases per occupation category.

946

947

Table 2 shows the percentage of men from each aptitude comparison group who were still in service as of December 1984. From that table it's apparent that the inadvertently enlisted men are just as likely to now be in the career force as are those in the appropriate higher-aptitude comparison group. Although retention rates are generally higher in the more demanding occupations (which are regarded as more desirable), no consistent aptitude-by-requirements interaction emerges.

A more sensitive index of career performance is paygrade achieved. Early paygrades are awarded pretty much in lock-step fashion (although there are considerable differences across services as to when the steps are taken), but promotions above E-4 are increasingly competitive. For those FY 1977 entrants in our aptitude comparison groups who were still in service in December 1984, Table 3 shows the proportion that had attained E-5 or better, and Table 4 shows the proportion that had attained E-6 or better. (Because of service differences in promotion policies, a floor effect in the Air Force E-6 data makes E-5 attainment more useful for examining performance within that service.)

These paygrade data are the first we have examined that show consistent mean differences favoring the higher-aptitude groups. Even here, the differences are not large. Analyses of the proportion of accessions still in service that attained a paygrade of E-5 or higher found no significant difference at $p<.01$ and only a single difference (for Marine Corps high school graduates in low requirement occupations) significant at $p<.05$. E-6 differences attain significance only (1) in the Army among high school graduates in occupations with low demands ($p<.01$) and (2) in the Navy for graduates in medium-demand occupations ($p<.05$) and nongraduates in high-demand occupations ($p<.01$). Moreover, it should be recognized that advancement is not based on on-the-job performance alone; individual scores on written tests for advancement are weighed and, in some cases, minimum scores on aptitude tests such as the AFQT are required as well. In this light, observed differences in advancement appear real, but surprisingly small.

Overall, we can make only a few generalizations at this point about the type of jobs in which low-aptitude individuals perform best. They are less likely than higher-aptitude men to become attrition cases within low- and medium-demand occupations, but there is no consistent pattern with regard to occupation type for the retention and promotion measures. What we can say about "when" low-aptitude recruits perform well is that they look quite acceptable when one looks at attrition or retention figures within an education/occupation group. They look less impressive when the analyst examines attainment of higher NCO ranks, but even here advancement appears to depend more on occupation type within a service than on aptitude category.

References

Greenberg, I. M. (1980). Report on mental standards for enlistment. In Department of Defense, Implementation of new Armed Services Vocational Aptitude Battery and actions to improve the enlistment standards process. Washington, DC: OASD (Manpower, Reserve Affairs, & Logistics).

Shields, J. L., & Grafton, F. C. (1983). A natural experiment: Analysis of an almost unselected Army population. Alexandria, VA: Army Research Institute.

Table 3

Percent of FY 1977 Male Accessions Still in Service as of December 1984 Who Had Achieved Paygrade E-5 or Higher, by Occupational Requirement Level, Service and Education, for Selected Aptitude Groups (N appears in parentheses)

Occupational Requirement	ARMY				NAVY				AIR FORCE				MARINE CORPS			
	HSG		NHS		HSG		NHS		HSG		NHS		HSG		NHS	
	IVC	IVB	IVA	IIIB	IVB	IVA	IVA	IIIB	IVB	IVA	IIIA	II	IVB	IVA	IVB	IVA
Low	85 (774)	83 (959)	79 (660)	79 (494)	70 (235)	68 (308)	63 (120)	67 (210)	--[a] (13)	74 (127)	-- (23)	83 (63)	81* (185)	89 (263)	-- (8)	79 (58)
Medium	80 (891)	83 (1,021)	74 (760)	77 (697)	73 (590)	76 (1,092)	66 (288)	72 (511)	73 (60)	82 (587)	85 (115)	87 (232)	81 (122)	84 (216)	-- (14)	82 (54)
High	82 (354)	83 (590)	82 (409)	84 (439)	76 (103)	79 (250)	76 (53)	81 (139)	-- (9)	75 (79)	-- (42)	86 (135)	-- (38)	-- (58)	-- (5)	-- (13)

*Significantly different from higher-aptitude comparison group at $p<.05$.
[a]Fewer than 50 cases per occupation category.

Table 4

Percent of FY 1977 Male Accessions Still in Service as of December 1984 Who Had Achieved Paygrade E-6 or Higher, by Occupational Requirement Level, Service and Education, for Selected Aptitude Groups (N appears in parentheses)

Occupational Requirement	ARMY				NAVY				AIR FORCE				MARINE CORPS			
	HSG		NHS		HSG		NHS		HSG		NHS		HSG		NHS	
	IVC	IVB	IVA	IIIB	IVB	IVA	IVA	IIIB	IVB	IVA	IIIA	II	IVB	IVA	IVB	IVA
Low	33** (774)	42 (959)	37 (660)	40 (494)	12 (235)	12 (308)	9 (120)	11 (210)	--[a] (13)	2 (127)	-- (23)	5 (63)	21 (185)	25 (263)	-- (8)	22 (58)
Medium	15 (891)	18 (1,021)	18 (760)	20 (697)	20* (590)	25 (1,092)	18 (288)	20 (511)	0 (60)	1 (587)	2 (115)	6 (232)	27 (122)	31 (216)	-- (14)	31 (54)
High	24 (354)	30 (590)	25 (409)	31 (439)	22 (103)	26 (250)	9** (53)	27 (139)	-- (9)	1 (79)	-- (42)	4 (135)	-- (38)	34 (58)	-- (5)	-- (13)

*Significantly different from higher-aptitude comparison group at $p<.05$.
**Significantly different from higher-aptitude comparison group at $p<.01$.
[a]Fewer than 50 cases per occupation category.

Enlistment and Utilization of High-Aptitude Recruits

Janice H. Laurence and Elizabeth F. Schneider
Human Resources Research Organization

High-aptitude personnel are an important element in the overall manning of our military forces. In fiscal year (FY) 1984 alone, over 123,000 persons scoring in AFQT Categories I and II (the upper 35 percent of the distribution) were accessed into the active enlisted forces.

Concerns over the cessation of banner recruiting times, the quality demands imposed by increasing military specialization and complex weaponry, and increased competition from the private sector for quality youth lead to questions as to the efficient utilization of high-aptitude military personnel.

Most studies addressing issues of manpower quality have focussed on persons scoring within the upper half of the ability distribution as a single entity. This research represents a departure from this tactic by examining the enlistment, assignment, and performance of individuals within the various aptitude levels traditionally labeled as "quality" personnel. More specifically, this paper compares the historical proportions enlisted, assignment patterns, and military performance of high-(AFQT Categories I and II) and average-aptitude (AFQT categories IIIA and IIIB) personnel.

Historical Enlistment Trends. Approximately five percent of the current 18 to 23 year old male youth population, is in AFQT Category I, 35 percent is in Category II, and 34 percent is in Category III (Department of Defense, 1982). As Table 1 shows, from the early 1950s through the present, the Defense Department has exceeded the national percentage within Categories I through III combined. However, DoD's proportion of very high-aptitude personnel (particularly within Category I) has not remained consistently better than the national average. Very high quality it seems, was easier to come by when the draft was in effect. Prior to the All-Volunteer Force (AVF), DoD was exceeding the national percentage of Category I male youth by as much as four percent.

Table 1

Percent Distribution of Male Recruits (All Services Combined) by AFQT Category, FY 1952-83

Percent Distribution of Male Recruits[a]							
Fiscal Year	Category I	Category II	Category III	Fiscal Year	Category I	Category II	Category III
1952	6.4	22.0	32.3	1973	3.7	30.1	52.1
1953	7.2	24.1	31.5	1974	3.0	32.3	54.5
1954	8.2	25.3	36.9	1975	3.5	34.0	56.3
1955	7.8	25.3	38.1	1976	3.9	33.9	51.7
1956	7.1	25.9	40.2	1977	4.3	28.2	39.6
1957	7.8	25.2	42.8	1978	3.6	27.3	42.1
1958	8.7	26.2	47.1	1979	3.0	23.6	41.8
1959	9.1	27.8	47.7	1980	2.8	23.8	41.6
1960	8.2	26.9	51.3	1981	2.6	30.2	47.8
1961	6.1	31.3	49.7	1982	3.1	33.4	49.4
1962	6.2	31.8	45.7	1983	3.7	36.7	50.1
1963	6.0	32.5	47.8	1984	3.8	36.8	51.1
1964	6.3	32.1	47.1				
1965	5.5	31.3	48.8				
1966	6.4	33.5	43.5				
1967	6.6	33.1	38.7				
1968	6.0	31.8	37.6				
1969	6.2	31.7	37.7				
1970	5.3	30.5	41.0				
1971	5.1	30.0	43.1				
1972	4.2	30.2	48.1				

All-Volunteer Force Transition[b]

Sources: Data for FYs 1952-70 were calculated from *Annual Reports of the Qualitative Distribution of Military Manpower*. Data for FYs 1971-83 were provided by the Defense Manpower Data Center.

[a]Male recruits include persons without prior military service who were inducted or enlisted and entered active duty (all Services combined) during the indicated fiscal year.

[b]The official end of the draft occurred on 30 June 1973. The drawdown began in July 1972, with the last draft call issued in December 1972.

With the advent of the AVF and continuing today, there are proportionally fewer very high-aptitude youth in the military than in the general population. The percentage of male recruits scoring in AFQT Category I has dropped to as low as 2.6 in FY 1981. In FY 1984 DoD's percentage was 3.8, still below that of the male youth population.

Assignment of High and Average Aptitude Recruits. The assignment patterns of FY 1979 through 1983 male high- and average-aptitude recruits across jobs of low, medium, and high skill requirements are presented below in Tables 2A-2D. Nine of the 10 DoD occupation codes were placed in the various skill requirement or job difficulty categories according to their mean ASVAB composite requirements (Sticht & Caylor, 1982) as follows:

Low	Medium	High
Infantry	Administrative	Electronics Repair
Service/Supply	Electrical/Mechanical Repair	Communications
	Craftsmen	Medical Corps
		Other Technical

The "Non-Occupational" category was omitted from the present analyses.

Overall, across all Services, there was an inverse relationship between the proportion of male recruits and AFQT category for low skill level jobs. That is, there were 17% Category Is, 27% Category IIs, 32% Category IIIAs, and 39% Category IIIBs in low difficulty jobs. This pattern held among education groups. This inverse relationship was also found among medium difficulty jobs, though the differences between the four aptitude groups were not as pronounced. For high skill requirement jobs, the relationship was direct. That is, the higher the aptitude category, the greater the proportion of recruits assigned within those occupational areas. The proportions of Category IIIB to Category I recruits within high cognitive skill requirement jobs were 18%, 27%, 33%, and 48%, respectively.

Interesting findings emerged when the data were examined across job skill requirements within AFQT categories. First, as one might expect, with increasing job skill demands, there was a higher proportion of Category I recruits, with 17% in low, 35% in medium, and 48% in high difficulty jobs. This pattern did not hold for any other aptitude group examined.

Most AFQT Category II males were assigned to medium skill jobs (40%), followed by high (33%) and then low (27%). Yet another pattern was shown among job difficulty levels for AFQT Categories IIIA and IIIB recruits. Most of these average-aptitude personnel served in medium skill requirement occupational areas followed by low and then high skill jobs.

Although the above comments are generalizations drawn across Services, Tables 2A-2D show that the Services differed somewhat in their assignment patterns for high- and average-aptitude personnel. Furthermore, there were dramatic Service differences in the proportions of recruits of various aptitude categories assigned among low, medium, and high skill occupations. Such Service differences, perhaps, can be attributed in large part to the variation in Service specific missions, job structure, and aptitude mixes. The Army and Marine Corps, for example, have more combat or low skill requirement jobs while the Navy and Air Force have a greater proportion of highly technical jobs.

Table 2A (ARMY)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

		Occupational Area Technical Skill Requirement[a]						
		Low		Medium		High		Total[b]
AFQT Category	Education	N	%	N	%	N	%	N
CAT I	HSG	3845	35.5	2146	19.8	4841	44.7	10832
	NHS/GED	258	34.5	171	22.9	319	42.6	748
	Total	4103	35.4	2317	20.0	5160	44.6	11580
CAT II	HSG	36322	40.5	21385	23.9	31891	35.6	89598
	NHS/GED	8946	44.4	5336	26.5	5848	29.1	20130
	Total	45268	41.3	26721	24.4	37739	34.4	109728
CAT IIIA	HSG	24875	43.2	16032	27.8	16710	29.0	57617
	NHS/GED	13394	48.8	7733	28.2	6334	23.1	27461
	Total	38269	45.0	23765	27.9	23044	27.1	85078
CAT IIIB	HSG	36128	43.6	28518	34.4	18264	22.0	82910
	NHS/GED	22635	52.2	12417	28.6	8298	19.1	43350
	Total	58763	46.5	40935	32.4	26562	21.0	126260

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

[b] Those not in job status have been excluded.

Table 2B (NAVY)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

		Occupational Area Technical Skill Requirement[a]						
		Low		Medium		High		Total[b]
AFQT Category	Education	N	%	N	%	N	%	N
CAT I	HSG	264	2.3	6398	55.8	4809	41.9	11471
	NHS/GED	68	7.3	296	31.9	565	60.8	929
	Total	332	2.7	6940	56.0	5374	43.3	12400
CAT II	HSG	3925	4.9	39475	49.1	36934	46.0	80334
	NHS/GED	1590	10.0	7662	48.2	6652	41.8	15904
	Total	5515	5.7	47137	49.0	43586	45.3	96238
CAT IIIA	HSG	3671	9.0	21455	52.5	15550	38.0	40875
	NHS/GED	2314	16.1	8328	58.1	3703	25.8	14345
	Total	5985	10.9	29783	54.1	19253	35.0	55021
CAT IIIB	HSG	6142	14.7	24880	59.4	10837	25.9	41859
	NHS/GED	2617	21.0	7687	61.7	2162	17.3	12466
	Total	8759	16.1	32567	60.0	12999	24.0	54325

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

[b] Those not in job status have been excluded.

863

Table 2C (MARINE CORPS)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

		Occupational Area Technical Skill Requirement[a]						
		Low		Medium		High		Total[b]
AFQT Category	Education	N	%	N	%	N	%	N
CAT I	HSG	854	28.8	971	32.7	1140	38.5	2965
	NHS/GED	116	39.2	176	59.5	89	30.1	296
	Total	970	29.7	1062	32.6	1229	37.7	3261
CAT II	HSG	12195	36.8	12140	36.6	8843	26.7	33178
	NHS/GED	3549	55.0	1825	28.3	1073	16.7	6447
	Total	15744	39.7	13965	35.2	9916	25.0	39625
CAT IIIA	HSG	11136	46.5	9009	37.6	3827	16.0	23972
	NHS/GED	4814	64.6	1714	23.0	923	12.4	7453
	Total	15950	50.8	10723	34.1	4752	15.1	31425
CAT IIIB	HSG	19860	57.0	11326	32.5	3705	10.6	34898
	NHS/GED	6735	73.8	1625	17.8	772	8.5	9132
	Total	26595	60.4	12951	29.4	4484	10.2	44030

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

[b] Those not in job status have been excluded.

953

Table 2D (AIR FORCE)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

		Occupational Area Technical Skill Requirement[a]						
		Low		Medium		High		Total[b]
AFQT Category	Education	N	%	N	%	N	%	N
CAT I	HSG	800	8.6	2771	29.7	5754	61.7	9325
	NHS/GED	81	12.7	248	39.0	307	48.3	636
	Total	881	8.8	3019	30.3	6061	60.8	9961
CAT II	HSG	14288	16.7	32357	37.8	38998	45.5	85643
	NHS/GED	1723	17.2	4876	48.7	3409	34.1	10008
	Total	16011	16.7	37233	38.2	42407	44.3	95651
CAT IIIA	HSG	11711	22.5	27021	51.8	13391	25.7	52123
	NHS/GED	1367	20.7	4015	60.7	1235	18.7	6617
	Total	13078	22.3	31036	52.8	14626	25.0	58740
CAT IIIB	HSG	14858	27.1	33231	60.6	6730	12.3	54820
	NHS/GED	978	23.6	2777	67.1	385	9.3	4140
	Total	15836	26.9	36008	61.1	7116	12.1	58960

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

[b] Those not in job status have been excluded.

954

Table 3A (ARMY)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

AFQT Category	Education	Occupational Area Technical Skill Requirement[a] Low	Medium	High
CAT I	HSG	7.1	4.7	3.1
	NHS/GED	12.8	5.5	7.5
	Total	7.4	4.9	3.4
CAT II	HSG	8.4	5.0	4.7
	NHS/GED	16.0	10.7	11.4
	Total	9.9	6.1	5.7
CAT IIIA	HSG	9.1	5.9	5.4
	NHS/GED	17.6	12.1	12.1
	Total	12.1	7.7	7.2
CAT IIIB	HSG	9.7	6.5	6.2
	NHS/GED	16.7	12.2	11.7
	Total	12.4	8.2	7.9

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Electrical/Mechanical Repair, and Craftsmen. High includes Administrative, Electronics Repair, Communication, Medical Corps, and other Technical.

Table 3B (NAVY)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

AFQT Category	Education	Occupational Area Technical Skill Requirement[a] Low	Medium	High
CAT I	HSG	2.3	1.7	.8
	NHS/GED	4.4	1.4	1.1
	Total	2.7	1.7	.8
CAT II	HSG	2.4	1.5	.8
	NHS/GED	4.2	2.0	2.0
	Total	2.9	1.6	.9
CAT IIIA	HSG	2.8	.9	1.2
	NHS/GED	4.1	2.0	3.0
	Total	3.3	1.2	1.5
CAT IIIB	HSG	2.2	.7	1.3
	NHS/GED	3.6	1.5	3.1
	Total	2.6	1.0	1.6

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

Table 3C (MARINE CORPS)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

AFQT Category	Education	Occupational Area Technical Skill Requirement[a] Low	Medium	High
CAT I	HSG	2.0	1.1	.4
	NHS/GED	.9	1.1	1.1
	Total	1.9	1.1	.5
CAT II	HSG	2.0	.8	.5
	NHS/GED	3.3	1.6	1.7
	Total	2.3	.9	.6
CAT IIIA	HSG	2.0	.8	1.2
	NHS/GED	3.6	1.4	1.7
	Total	2.5	.9	1.3
CAT IIIB	HSG	1.9	.8	1.0
	NHS/GED	3.3	2.2	1.3
	Total	2.3	1.0	1.0

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

Table 3D (AIR FORCE)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

AFQT Category	Education	Occupational Area Technical Skill Requirement[a] Low	Medium	High
CAT I	HSG	3.3	2.6	1.0
	NHS/GED	11.1	8.1	5.2
	Total	4.0	3.0	1.2
CAT II	HSG	2.9	3.2	1.6
	NHS/GED	12.8	9.9	7.6
	Total	4.0	4.1	2.0
CAT IIIA	HSG	3.3	3.7	2.3
	NHS/GED	12.4	10.9	8.4
	Total	4.0	4.7	2.8
CAT IIIB	HSG	3.7	4.1	2.7
	NHS/GED	17.7	12.5	10.1
	Total	4.6	4.8	3.1

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

account for large numbers of "attritees"). The significant differences in attrition among aptitude categories most likely resulted from the relatively greater attrition differences between extreme aptitude categories (i.e., between AFQT Category I and IIIB).

Not only was there a lack of statistical significance between attrition and aptitude category within the various job skill demand levels but inspection for "non-significant" trends showed little as well. Similar to the Army, Air Force data showed an AFQT category-attrition rate inverse linear relationship. Statistical significance, however, was found only among nongraduates. Again, though χ^2s were computed across all four aptitude categories it appears that the differences were mainly between the extremes--Categories I and IIIB.

Concluding Note. Since the study of high-aptitude military personnel is yet in its beginning stages, extensive generalizations cannot (and should not) be drawn from these analyses. Despite this caveat, the assignment differences and attrition rates of high- and average-aptitude personnel among jobs of various skill requirements have been an informative first step in the investigation of the utilization of high-aptitude personnel. Service differences in assignment patterns and attrition rates within job difficulty levels abound. Regardless of such differences in the Services and their job requirements, high- and average-aptitude personnel are assigned differentially within the various skill demand occupational areas. Furthermore, for high- and average-aptitude recruits there are some tendencies toward higher attrition rates in low skill jobs. A clearer picture of the relationship between job difficulty or complexity may emerge by using a taxononomy independent of Service selection and classification standards and the broad DoD occupational area categorizations. These and further analyses may have important implications for a more efficient distribution and matching of available talent within jobs, and overall more effective use of limited resources.

References:

Department of Defense (1982, March). Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics).

Sticht, T. G., & Caylor, J. S. (1982, June). Evaluation of the Literacy Assessment Battery (LAB) as a predictor of success in the Armed Services (FR-ETSD-82-10). Alexandria, VA: Human Resources Research Organization.

Mapping Predictors to Criterion Space: Overview

Norman G. Peterson
Personnel Decisions Research Institute

Introduction

Our applied problem is to expand the presently measured predictor space for the ultimate purpose of accurately selecting persons for the U.S. Army and appropriately classifying those persons into jobs or Military Occupational Specialities (MOS). In this paper, I describe the strategy we have adopted, the thinking behind the strategy, and some of the progress that has been made following our strategy. A fuller description can be found in Peterson, 1985.

As you all know, the U.S. Army presently has a lot of jobs and hires, almost exclusively, inexperienced and untrained persons to fill those jobs. One implication of these obvious facts is that a highly varied set of individual differences' variables must be put into use to stand a reasonable chance of improving the present level of accuracy of predicting training performance, job performance, and attrition/retention in a substantial proportion, if not all, of those jobs. Much less obvious is the particular content of that set of individual differences variables, and the way the set should be developed and organized; or put another way, how the predictors should be mapped onto the criterion space.

Theoretical Approach

We have approached this problem by adopting a construct-oriented strategy of predictor development, but with a healthy leavening from the content-oriented strategy. Essentially, we endeavored to build up a model of predictor space by (a) identifying the major, relatively independent domains or types of individual differences' constructs that existed; (b) selecting measures of constructs within each domain that met a number of psychometric and pragmatic criteria, and (c) further selecting those constructs that appeared to be the "best bets" for incrementing (over present predictors) the prediction of the set of criteria of concern (i.e., training/job performance and attrition/retention in Army jobs). Ideally, the model would, we hoped, lead to the selection of a finite set of relatively independent predictor constructs that were also relatively independent of present predictors and maximally related to the criteria of interest. If these conditions were met, then the resulting set of measures would predict all or most of the criteria, yet possess enough heterogeneity to yield powerful, efficient classification of persons into different jobs. The development of such a model also had the virtue that it could be at least partially "tested" at many points during the research effort, and not just at the end, when all the predictor and criterion data are in. For example, we could examine the covariance of newly developed measures with one another and with the present predictors, notably the ASVAB. If the new measures were not relatively independent of ASVAB and measures from other domains as predicted by the model, then we could take steps to correct that. Also, by constructing such a visible model, we thought that modifications and improvements could be much more straightforwardly implemented.

Figure 1 presents an illustrative, construct-oriented model and is presented in order to represent the model in abstract. Note that both the criterion and predictor space are depicted. A great deal of the work of Project A is devoted to describing and defining the job performance criterion and we, on the predictor side, have made much use of the information coming from those efforts.

If this illustrative model were to be developed and tested with data, then the network of relationships on the predictor side, the criterion side, and between the two could be confirmed, disconfirmed, and/or modified. It goes without saying, but I will say it anyway, that the development of such models must be done very carefully and conservatively, and subjected frequently to reality testing. We have kept this firmly in mind, Note, however, that the possession of such a model enables one to state fairly clearly why such a predictor is being researched, and to check quickly, at least rationally, whether or not the addition of a predictor is likely to improve prediction.

Finally, the model is depicted as a matrix with a hierarchical arrangement of both the rows and columns. We have found it very useful to employ this hierarchical notion, since it allows us to think in terms of appropriate levels of specificity for a particular problem as we do the research, or for future applications of measures.

We began our research with a general kind of model, very much like the one presented in Peterson and Bownas (1982). That is, we conceived of the predictor space as divided into several domains with major, relatively independent constructs falling into each domain. At this early point in the research, we were most concerned with thinking about the predictor space in a way guided by past research that would also provide "handles," if you will, for us to approach our particular applied problem. We formed

		Criteria							
		Training Performance			Job Task Performance		Attrition/ Retention		
Predictors		Pass/ Fail	Test Grades	Atten- dance	Common Tasks	Specific Tasks	Finish Term	Reen- list	Early Discharge
Cognitive	Verbal	H*	H	L	H	H	L	L	L
	Numerical	H	H	.	.	.	.		
	Spatial								
Psychomotor	Precision								
	Coordination								
	Dexterity								
Temperament	Dependability								
	Dominance								
	Sociability								
Interests	Realistic								
	Artistic								
	Social	.	.	.	H	H	H	L	L

FIGURE 1. Illustrative Construct-Oriented Model

*Denotes expected strength of relationship, High, Medium, Low.

three domain teams to be responsible for broad pieces of this predictor space model, to wit: a "non-cognitive" team for temperament, biographical data, and vocational interest variables; a "cognitive" team for cognitive and perceptual variables; and a "psychomotor" team for psychomotor variables.

Literature Review. The domain teams began with a large-scale literature review. Within each area, the teams carried out essentially the same steps. These were: 1) compile an exhaustive list of possibly relevant reports, articles, books or other sources; 2) review each source and determine its relevancy for the project by examining the title and abstract (or other brief review); 3) obtain the sources identified as relevant in the second step; and 4) for relevant materials, carry out a thorough review and transfer relevant information onto special review forms developed for the project.

Within the first step, several activities were carried out to insure as comprehensive a list as possible. Several computerized searches of relevant data bases were done. In addition to the computerized searches, we obtained reference lists from recognized experts in each of the areas, emphasizing the most recent research in the field. We also obtained several annotated bibliographies from military research laboratories. Finally, we scanned the last several years' editions of research journals that are frequently used in each ability area as well as more general sources such as textbooks, handbooks, and appropriate chapters in the Annual Review of Psychology.

The vast majority of the sources identified as described above were not relevant to our purpose. These non-relevant sources were weeded out in Step 2. After obtaining the relevant sources, these were reviewed and two forms were completed for each source: an Article Review form and a Predictor Review form (several of the latter form could be completed for each source.) These forms were designed to capture, in a standard format, the essential information from the reviewed sources, which varied considerably in their organization and reporting styles. The output of the literature search, in the form of the completed review forms and copies of the actual sources, served as input to several later steps.

Expert Judgments. One of these steps was the identification of a set of predictor constructs that met a number of psychometric and practical criteria. There were twelve such criteria used to evaluate constructs, like reliability, criterion-related validity, robustness and ease of administration procedures, etc. At least two researchers evaluated each construct on these twelve factors, using five point scales, and these evaluations guided the selection of 53 predictor constructs.

Definitions of these selected constructs were written and descriptive materials (psychometric data, validity evidence, and illustrative items) were prepared. These materials were used in an expert judgment process wherein 35 experienced personnel and research psychologists estimated the "true validity" of each of the 53 predictor constructs for each of 72 Army enlisted criteria. These 72 criterion descriptions were prepared by Project A researchers who were focusing on describing the job performance of Army enlisted ranks. (See Wing, Peterson, and Hoffman, 1984, for a complete description of this expert judgment process.)

These expert judgments proved to be highly reliable (the reliability of the pooled raters' estimates of validity of each construct for each criterion was over .90), and factor analysis of their ratings provided our first model of the predictor space. Figure 2 shows that model. This

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension 5. Reading Comprehension 16. Ideational Fluency 18. Analogical Reasoning 21. Omnibus Intelligence/Aptitude 22. Word Fluency	A. Verbal Ability/ General Intelligence	
4. Word Problems 8. Inductive Reasoning: Concept Formation 10. Deductive Logic	B. Reasoning	
2. Numerical Computation 3. Use of Formula/Number Problems	C. Number Ability	COGNITIVE ABILITIES
12. Perceptual Speed and Accuracy	H. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory 17. Follow Directions	J. Memory	
19. Figural Reasoning 23. Verbal and Figural Closure	F. Closure	
6. Two-dimensional Mental Rotation 7. Three-dimensional Mental Rotation 9. Spatial Visualization 11. Field Dependence (Negative) 15. Place Memory (Visual Memory) 20. Spatial Scanning	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental Information Processing	INFORMATION PROCESSING
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests 51. Artistic Interests (Negative)	M. Realistic vs. Artistic Interests	
28. Control Precision 29. Rate Control 32. Arm-hand Steadiness 34. Aiming	I. Steadiness/Precision	
27. Multilimb Coordination 35. Speed of Arm Movement	D. Coordination	PSYCHOMOTOR
30. Manual Dexterity 31. Finger Dexterity 33. Wrist-Finger Speed	K. Dexterity	
39. Sociability 52. Social Interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising Interests	R. Enterprising Interest	
36. Involvement in Athletics and Physical Conditioning 37. Energy Level	T. Athletic Abilities/Energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/Self-esteem	
40. Traditional Values 43. Conscientiousness 46. Non-delinquency 53. Conventional Interests	N. Traditional Values/Conventionality/Non-delinquency	
44. Locus of Control 47. Work Orientation	O. Work Orientation/Locus of Control	MOTIVATION/ STABILITY
38. Cooperativeness 45. Emotional Stability	P. Cooperation/Emotional Stability	

FIGURE 2. Hierarchical Map of Predictor Space

PILOT TRIAL BATTERY	CLUSTERS	FACTORS
ASVAB	A. Verbal Ability/ General Intelligence	
Reasoning 1 and 2	B. Reasoning	
Number Memory (c)	C. Number Ability	COGNITIVE ABILITIES
Perceptual Speed and Accuracy (c) Target Identification (c)	H. Perceptual Speed and Accuracy	
AVOICE	U. Investigative Interests	
Short Term Memory (c)	J. Memory	
Reasoning 1 and 2	F. Closure	
Assembling Objects Object Rotation Shapes Mazes Path Orientation 1, 2, and 3	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
Simple Reaction Time (c) Choice Reaction Time (c)	G. Mental Information Processing	INFORMATION PROCESSING
ASVAB	L. Mechanical Comprehension	MECHANICAL
AVOICE	M. Realistic vs. Artistic Interests	
Target Tracking 1 (c) Target Shoot (c)	I. Steadiness/Precision	
Target Tracking 2 (c) Target Shoot (c)	D. Coordination	PSYCHOMOTOR
--	K. Dexterity	
ABLE/AVOICE	Q. Sociability	SOCIAL SKILLS
AVOICE	R. Enterprising Interest	
ABLE	T. Athletic Abilities/Energy	VIGOR
ABLE	S. Dominance/Self-esteem	
ABLE	N. Traditional Values/Conventionality/Non-delinquency	MOTIVATION/ STABILITY
ABLE	O. Work Orientation/Locus of Control	
ABLE	P. Cooperation/Emotional Stability	
Cannon Shoot (c)	Movement Judgment	

(c) = Computerized Measures

FIGURE 3. Pilot Trial Battery Measures of the Modeled Predictor Space

model represents the predictor structure in terms of their covariances with each other based on their judged validity relationships to dimensions of Army enlisted criteria.

Test Construction. Figure 2 served as a blueprint of sorts for our test construction efforts. The three domain teams set about writing tests and inventories to measure the constructs shown there. We went through a fairly extensive process of writing (or, in the case of computerized tests, programming) instruments, trying them out at Army sites (MEPS and/or Army forts), then revising the instruments based on the tryout results. After about four such iterations (at Minneapolis MEPS, Fts. Carson, Campbell, and Lewis), we possessed a set of instruments collectively labeled the Pilot Trial Battery. That set of measures is shown in Figure 3.

Note that the measures are slotted into the cluster and factor space, insuring that we adequately operationalized the model. Note also that one measure, "cannon shoot", is included and it measures Movement Judgment, a variable that was not originally included. It was added because it seemed to be a variable that was important for a variety of combat arms MOS, but had escaped our notice because of a dearth of research on such a variable.

This Pilot Trial Battery consumed approximately six and one-half hours of testing time and the entire battery was administered to a sample of about 250 soldiers at Ft. Knox. Test-retest data were also collected. Analyses of these data were used to further revise the measures and to reduce the battery in size so that it could be administered in four hours. The reduction in the size of the battery was accomplished by deleting some tests entirely and by deleting items from other tests. (The tests deleted were Reasoning 2, Shapes, Path, and Orientation 1.) The existence of the predictor model proved especially helpful to those of us faced with the hard decision of deleting tests and items. The impact of various decisions in terms of coverage of the "predictor space" could readily be seen and, along with the tryout data, empirically evaluated.

This revised and reduced battery was labeled the Trial Battery and is presently being administered to a large sample (N=11,000) of soldiers in the U.S. and Europe in a concurrent validity study. In terms of testing time, 34% of the battery is devoted to the computerized perceptual/psychomotor measures, 50% to cognitive paper-and-pencil measures, and 16% to non-cognitive, paper-and-pencil inventories. Once the concurrent validity data are in hand, we will be able to make some fairly definitive tests of our model--in terms of its factorial structure, validity, and classification efficiency.

References

Peterson, N. G., & Bownas, D. A. (1982). Skill, task structure, and performance acquisition. In Marvin D. Dunnette and Edwin A. Fleishman (Eds.), Human performance and productivity (Vol. I). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Peterson, N. G. (1985). Overall strategy and methods for expanding the measured predictor space. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Wing, H., Peterson, N. G., & Hoffman, R. E. (1984). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto.

Using Microcomputers for Assessment: Practical Problems and Solutions

Rodney L. Rosse and Norman Peterson
Personnel Decisions Research Institute

Introduction

"History repeats itself" is an adage that probably does not apply to the advances of microprocessor developments. Given the frantic rate of development, it is difficult to imagine that circumstances could ever again occur in just the way that they did at the outset of this effort in the Fall of 1983. It would seem, however, that any 1986 project might be enhanced by consideration of both the occasional wisdom and sometime folly of our beginning efforts.

Initially, even the goals to be accomplished were far from obvious and may have remained beyond our vision except for the valuable help obtained through visits to several research centers doing advanced work in computerized testing: (1) Air Force Human Resources Laboratory at Brooks Air Force Base, Texas, (2) Army Research Institute Field Unit at Fort Rucker, Alabama, (3) Naval Aerospace Medical Research Laboratory, Pensacola, Florida, and (4) Army Research Institute Field Unit at Fort Knox, Kentucky. Experimental testing projects using computers at these sites had already produced impressive developments which stimulated the ideas of the project at hand and have continued to influence our work.

In this paper, we focus primarily on the process we followed and some problems we encountered in hardware and software acquisition and development for the purpose of developing new predictor tests of abilities that could best be administered via microprocessors.

Hardware Acquisition and Development

Much of the detail of the planned products was yet to evolve at the point of acquisition of the first six machines so that we had to focus upon more general objectives. It was clear that we wished to accomplish several things which were either difficult or impossible to accomplish with paper-and-pencil testing. Specifically, we required the ability to have a very high degree of precision in stimulus presentation and a high degree of control of respondent behavior. Dependent variables were specifically expected to include precision in timing of stimulus presentation and response speed.

Microprocessor. The choice of which microprocessor to use for the preliminary development was not obvious. The arrays of available microcomputer devices were, at the time, in transition from earlier machines which used the first popular microprocessor chips (i.e., 8080 or Z-80) into a newer variety of options created by the influence of IBM's entry into the market with their "PC" employing the newer 8088, 8086-7 chips. With the newer machines came more flexible operating systems (e.g., DOS 1 or DOS 2).

A computer designed for portable use was deemed to be a highly desirable characteristic because the machines were to be frequently disassembled, carried to new locations, and reassembled by non-technical personnel. Such portable machines had been available only briefly so that little reported experience with them was available.

We acquired six machines made by Compaq (TM) which appeared to suit the need. They were among the "newer" types of machines which used a variation of the MS-DOS operating system. They were equipped with standard game adapters which permitted the analog inputs from "off-the-shelf" joysticks and boolean input from game button switches.

The choice was specifically made to avoid using color in the visual displays for at least two reasons: (1) the certainty of individual differences in color vision among military recruits, and (2) dread of the prospects of attempting to calibrate video colors for standardization of presentation. Accordingly, we precluded the possibility of directly investigating the value of stimulus effects in color presentation.

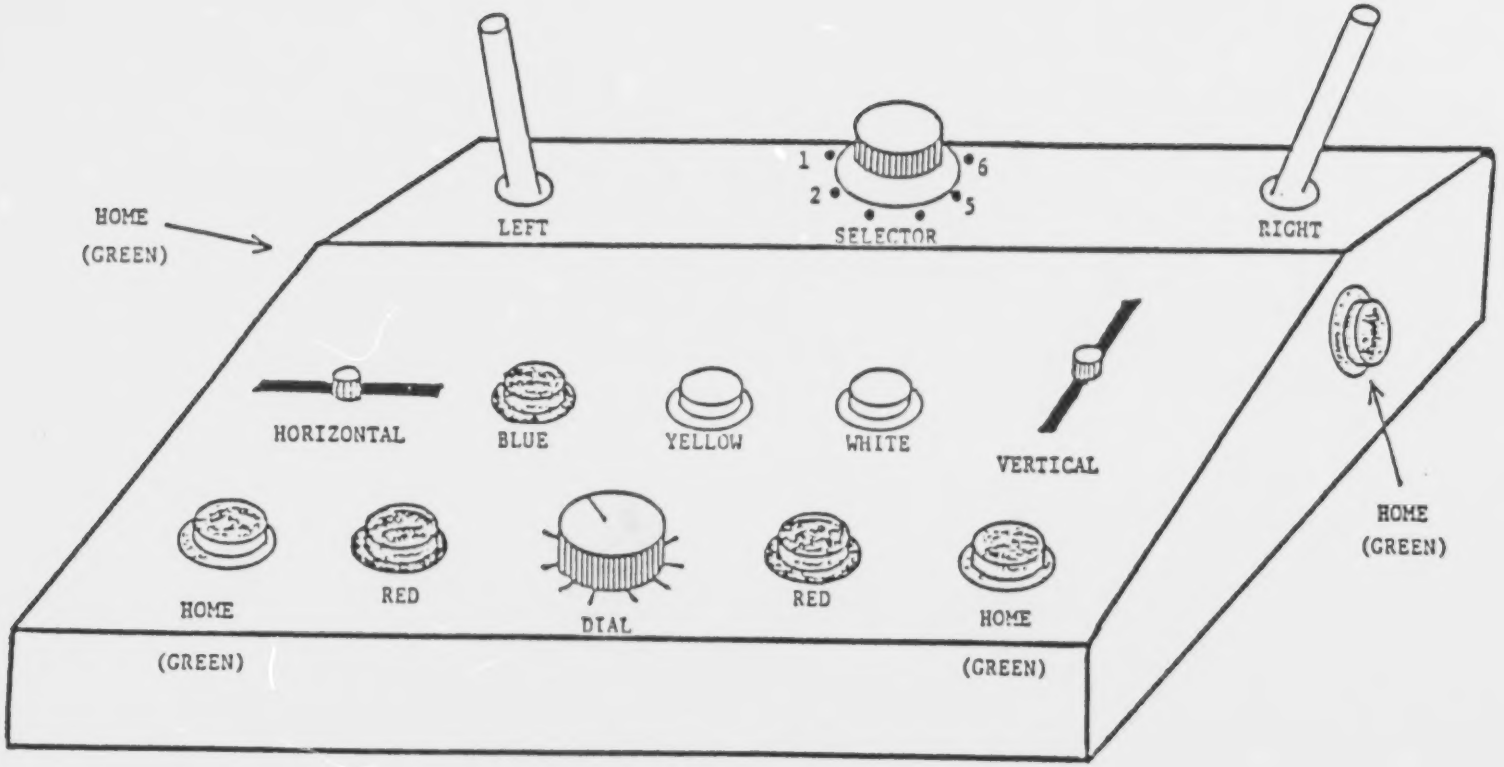
The graphics capability of the Compaq microcomputer proved to be minimally acceptable for the applications which were to come. In graphics mode, the pixels (or dots) on the screen are organized into 200 rows and 640 columns. More recently, several computers of the "personal" computer type are offering 400 rows with 640 columns which should provide improved resolution.

Very accurate timing of events occurring in the testing process was essential. Initially, timing was accomplished by two means: (1) accessing the calendar clock that is available in any machine which uses MS-DOS (or the variations of MS-DOS that are sold under computer tradenames), and (2) use of calibrated software loops. Without delving too far into technical details, those two options eventually presented some difficulties because of time consumption in the process of obtaining the time. For instance, the computer CPU often had to be tied up with timing events when other work required being done in the timed interval.

A wonderful solution to the timing problem eventually presented itself in what the computer people call a "real-time-clock" which can be added to the "IBM-type" microcomputers for as little as $50. Operating on a small battery it maintains the correct date and time even when the computer is turned off. With appropriate software, the "real-time-clock" device allows the timing of events accurately to the nearest 1/1000-th of a second with negligible loss of computer time in the reading. (The sub-program used in our projects will read the time in approximately 1/3000-th of a second.)

Peripheral Devices for Response Acquisition: Response Pedestal. The initial choices in the hardware configuration for a "testing station" proved satisfactory for the "stimulus side", i.e., the controlled presentation to the subject. The standard keyboard and the "off-the-shelf" joysticks were hopelessly inadequate for the "response side." Computer keyboards leave much to be desired as response acquisition devices--particularly when response latency is a variable of interest. Preliminary trials using, say, the "D" and "L" keys of the keyboard for "true" and "false" responses to items was troublesome with naive subjects. Intricate training was required to avoid individual differences arising from differential experience with keyboards. Moreover, the software had to be contrived so as to flash a warning when a respondent accidentally pressed any other key. The "off-the-shelf" joysticks were sadly lacking in precision of construction such that the score of a respondent depended heavily upon which joystick she/he was using.

We came up with a plan for a "response pedestal" which consisted of readily available electronic parts. A prototype of the device was obtained from a local engineer. (See Figure 1.) It had two joysticks, a horizontal and a vertical sliding adjuster, and a dial. The two joysticks allowed either left or right hand usage. The sliding adjusters permitted two-handed coordin-



875

967

FIGURE 1. Custom-designed response pedestal

968

ation tasks. The dial permitted respondent selections in a manner similar to the now popular "mouse" devices that are sold for "personal computers."

The response pedestal had nine button-switches, each of which was to be used for a particular purpose. Three buttons (BLUE, YELLOW, and WHITE) were located near the center of the pedestal and were used for registering up to 3-choice alternatives. Also near the center were two buttons (RED) which were mostly used to allow the respondent to step through frames of instructions and, for some tests, to "fire" a "weapon" represented in graphics on the screen.

Of notable interest was the placement of the button-switches which were called "HOME" with respect to the positions of other buttons used to register a differential response. The "HOME" buttons required the respondent's hands to be in the position of depressing all four of the "HOME" buttons prior to presentation of an item to which (s)he would respond. This, it is believed, offered advantages of control of attention and control of hand position for measurement of response latency. Using appropriately developed software, we were able to measure total response time but also to break it down into two parts: (1) "decision time" which is defined as the interval between onset of stimulus and release of the "HOME" keys, and (2) "movement" time which is the subsequent interval to the registering of a response. It was possible, where of interest, to even tell quite reliably whether the respondent used a left hand or a right hand to respond since (s)he almost invariably would release the "HOME" buttons on the side of the preferred hand first.

The rotary switch marked "SELECTOR" in Figure 1 was an inconvenience that was required by our initial choice of "game-adapter" for reading analog input. The game adapter initially chosen allowed only four inputs and the response pedestal had seven analog outputs: 2 inputs for each of two joysticks, two sliding adjusters, and one rotary adjuster called the "DIAL." The "SELECTOR" was used to select which analog devices were to be operative for a particular test item. The final design for the response pedestal included a game-adapter with the capability of eight analog inputs and the "SELECTOR" switch was happily omitted.

Joysticks. Perhaps the greatest difficulty regarding the response pedestal design arose from the initial choice of joystick mechanisms. We soon discovered that joystick design is a complicated and, in this case, a somewhat controversial issue. Variations in tension or movement can defeat the goal of standardized testing. While "high-fidelity" joystick devices are available, they can cost thousands of dollars apiece which was prohibitively expensive in the quantities that were to be required for this project. The first joystick mechanism that was used in the response pedestals was an improvement over the initial "off-the-shelf" toys that predated the pedestals. It had no springs whatsoever so that spring tension would not be an issue. It had a small, light weight handle so that enthusiastic respondents could not gain sufficient leverage to break the mechanism. It was inexpensive.

Unfortunately, this joystick had a "wimpy" feeling which was greatly lacking in "face-validity" (or, as Hilda Wing dubbed it, "fist-validity") from the Army's point of view. It was felt that the joystick was so much like a toy that it would not command respect of the respondents. It was the contention of a minority of us that our "wimpy" device had "construct fidelity" in that it would do a perfectly adequate job of testing the constructs that were targeted.

The joystick mechanism had to be changed. Joysticks of every conceivable variety and type of use were considered. We learned about viscous dampening,

friction, tension, and even something called "stiction." Ultimately, a joystick device was fashioned with a light spring for centering and a sturdy handle with a bicycle handle-grip. It had sufficient "fist-validity" to be accepted by all (or almost all) and it was sufficiently precise in design that we were unable to detect any appreciable "machine" effects in fairly extensive testing.

Software Development

We wish to turn attention now to the issues of software development. There were no "package programs" available to administer computerized tests. The selection of strategy for organizing and programming the needed software was to fall upon ourselves. We had three general, operational objectives in mind for the software to be produced: (1) as far as possible, it should be transportable to other microprocessors; (2) it should require as little intervention as possible from a test administrator in the process of presenting the tests to subjects and storing the data; and, (3) it should enhance the "standardization" of testing by adjusting for hardware differences across computers and response pedestals.

Primary Language. We chose to prepare the bulk of the software using the Pascal language as implemented by Microsoft, Inc. There were certain advantages to this in that Pascal is a common language and it is implemented using a compiler that permits modularized development and software libraries. As computer languages go, Pascal is relatively easy for others to read and it can be implemented on a variety of computers.

Some processes, mostly those which are specific to the hardware configuration had to be written in IBM-PC assembly language. Examples of these include the interpretation of the response pedestal inputs, reading of the real-time-clock registers, calibrated timing loops, and specialized graphics and screen manipulation routines. For each of these identified functions, a Pascal-callable "primitive" routine with a unitary purpose was written in assembly language. The functions were designed to be simple and unitary in purpose so as to be easily reproducible for other machines.

Strategy. The overall strategy of the software development is worth discussing. It quickly became clear that the direct programming of every item in every test by one person was not going to be very successful either in terms of time constraints nor in terms of quality of product. For the sake of making it possible for each researcher to contribute his/her judgment and effort to the project, it was necessary to plan so as to take the "programmer" out of the step between conception and product as much as possible.

The testing software modules were designed as "command processors" which interpreted relatively simple, problem oriented commands. These were organized in ordinary text written by the various researchers using word processors. Many of the commands were common across all tests. For instance, there were commands that permitted writing of specified text to "windows" on the screen and controlling the screen attributes (brightness, background shade, etc). A command could hold a display on the screen for a period of time (measured to 1/100-th second accuracy). There were commands which caused the program to wait for the respondent to push a particular button on the pedestal. Some of the commands were specific to particular item types. These commands were selected and programmed according to the needs of a particular test type. For each item type, we would decide upon the relevant stimulus

properties to vary and build a command that would allow the item writer to quickly construct a set of commands for items which she/he could then inspect on the screen.

Thus, entire tests were constructed and experimentally manipulated by psychologists who could not program a computer.

The strategies for developing commands have evolved and improved over the period of development. Eventually, the commands became almost "language-like" with syntax forms analogous to some of the common statistical packages like SPSS or SAS that are available on "main-frame" computers.

Hardware Testing and Calibration. One of the most useful software developments relates to the testing and calibration of the hardware, necessary for purposes of standardization. A complete hardware testing and calibration process can be undertaken by test monitors each time a machine is powered up. It checks the timing devices and screen distortion, and calibrates the analog devices (joysticks, sliding adjusters, dial) so that measurement of movement will be the same across machines. It also permits the software adjustment of the height to width ratio of the screen display so that circles do not become ovals or, more importantly, the relative speed of moving displays remains under control regardless of vertical or horizontal travel.

Concluding Remarks

In the end, we were able to put together a portable, complete testing session lasting approximately 1-1/2 hours where very naive respondents can complete the test with little or no intervention from a test monitor. The data is automatically stored and "backed-up" on diskettes in a form readily transferrable to a "main-frame" for analysis. Except for occasional calibration or contingencies, the test monitor needs only to turn the computers on and put the respondents in front of them.

Finally, and perhaps most gratifying, we have found that the soldiers tested via this method have generally preferred computerized testing to paper-and-pencil testing. We have not gathered hard data on this aspect, but base our conclusions on observation of the soldiers while taking the battery and their comments to us after completing the battery. Perhaps this is due to novelty alone, but we feel it may also be due to the nature of the tests themselves plus the fact that the soldier, in large part, is in control of the testing process her/himself. They control the pacing of instructions for the tests and, for some tests, the pacing of item presentation. No administrator tells them when to begin and when to stop, and they are not in "lock step" with a larger group. We view this state of affairs as highly desirable for personnel selection testing.

Computerized Assessment of Perceptual and Psychomotor Abilities

Jeffrey J. McHenry and Jody L. Toquam

Personnel Decisions Research Institute

One of the main goals of the Army Research Institute's (ARI's) Project A is to develop new predictor measures to supplement the Armed Services Vocational Aptitude Battery (ASVAB). In this paper, we describe 10 new computerized perceptual and psychomotor predictor tests that were pilot tested last fall and are currently being validated in a large-scale concurrent validation study.

The Computer Battery

Toquam, Dunnette, Corpe, and Houston, (1985) have described the procedures used to identify target constructs for cognitive-perceptual predictor test development, and to determine which of these constructs would be measured via paper-and-pencil tests and which would be measured via computer. Following a similar procedure, members of the Project A research team working in the psychomotor ability domain identified two psychomotor ability constructs for predictor test development. Since measurement of both of these constructs required that subjects be presented with a moving stimulus object, it was decided that all psychomotor tests would be presented on the computer.

In total, computer tests were developed for seven constructs (i.e., five cognitive-perceptual ability constructs and two psychomotor ability constructs). To measure these seven constructs, 10 new computer tests were developed. The constructs and tests are listed in Table 1. As Table 1 shows, two tests each were developed to assess reaction time, perceptual speed and accuracy, and precision/steadiness, while one test each was developed to assess the remaining four constructs. (Complete descriptions of each test are available from the authors upon request.)

TABLE 1

Target Constructs and Computer Tests

Target Construct	Definition	Test(s)
Reaction Time	The ability to detect a simple stimulus quickly	Simple Reaction Time Choice Reaction Time
Perceptual Speed and Accuracy	The ability to compare two stimuli and to determine quickly and accurately whether they are the same or different	Perceptual Speed and Accuracy Target Identification
Memory	The ability to encode and store information, and then retrieve that information quickly and accurately	Short Term Memory
Number Facility	The ability to perform simple numerical operations (e.g., addition, subtraction, multiplication, division) quickly and accurately	Number Memory
Movement Judgment	The ability to judge the movement speed and direction of an object and to determine when (or whether) that object will reach a given point in space	Cannon Shoot
Multilimb Coordination	The ability to coordinate the use of two or more limbs (e.g., two hands, two feet, a hand and a foot, etc.) to perform a task	Target Tracking 2
Steadiness/ Precision	The ability to make fine coordinated movements in response to a moving stimulus object	Target Tracking 1 Target Shoot

Pilot Testing

During test development, several pilot tests of portions of the computer battery were conducted at Ft. Carson, Ft. Lewis, and the Minneapolis Military Enlistment Processing Station. A more extensive pilot test of the entire battery was then conducted last fall at Ft. Knox.

The purpose of the pilot testing was to ensure that the tests satisfied three criteria for administration in the Project A concurrent validation study. First, we wanted to ensure that the 10 tests were reliable. Second, we wanted to make certain that the tests did not overlap greatly with the ASVAB. Finally, we wanted to ensure that the computer tests themselves are not highly intercorrelated, since our goal is to measure seven distinct ability constructs with these 10 tests.

Method

Subjects

Subjects included 256 first-term Army enlisted personnel stationed at Ft. Knox. Subjects were drawn from a wide range of MOS. All subjects had been in the service between one and two years at the time of testing.

Procedure

When subjects arrived in the computer testing room, they were asked to take a seat at a testing station. They were told that the computer tests were self-administering so they could work at their own pace. They were instructed to read the instructions carefully, ask questions if they encountered any problems, and try their hardest.

Two weeks later, 121 of the subjects returned for retesting. They were given the same instructions that they had received two weeks earlier and asked to complete the entire computer battery a second time.

Results

Scoring

Responses on computer tests may be used to compute numerous scores. For example, responses to Perceptual Speed and Accuracy items, may be summarized using average decision time, average movement time and average total response time across all items or across only those items in which the subject responds correctly. The average response for each of these may consist of the mean, the median or a trimmed mean computed by deleting the fastest and slowest response times. Other dependent measures derived from this test include the slope and intercept which are computed by regressing the subject's response time against some specified item parameter such as item length. Finally, percent correct can be used as a dependent measure for each subject.

In total, for the 10 tests, 168 different test scores were computed. Preliminary analyses of the reliability of each score and the intercorrelations among the various scores within each test were used to reduce this list to 19 test scores (see Table 2). These 19 scores received more extensive analyses.

Reliability

Table 2 contains the split-half and test-retest reliability for each test score. The majority of split-half reliabilities exceeded .80, and only two are less than .70. As expected, the test-retest reliabilities are lower than the split-half reliabilities. Five test scores have test-retest reliabilities less than .55. In general, those test scores with low test-retest

TABLE 2

Characteristics of the 19 Computer Test Scores

Test Score	Reliability r_{sh}	Reliability r_{tt}	Overlap with ASVAB SMC	Overlap with ASVAB Uniqueness
COGNITIVE-PERCEPTUAL TESTS				
Simple Reaction Time - Mean Rt	.90	.37	.07	.83
Choice Reaction Time - Mean Rt	.89	.56	.09	.80
Perc Speed & Acc - Pct Correct	.83	.59	.14	.69
Perc Speed & Acc - Mean RT	.96	.65	.06	.90
Perc Speed & Acc - Slope	.88	.67	.09	.79
Perc Speed & Acc - Intercept	.74	.55	.11	.63
Target Ident - Pct Correct	.84	.19	.05	.79
Target Ident - Mean RT	.96	.67	.16	.80
Short Term Memory - Pct Correct	.72	.34	.10	.62
Short Term Memory - Mean RT	.94	.78	.06	.88
Short Term Memory - Slope	.52	.47	.01	.51
Short Term Memory - Intercept	.84	.74	.11	.73
Number Memory - Pct Correct	.63	.53	.40	.23
Number Memory - Mean Oper RT	.95	.88	.33	.62
Cannon Shoot - Time Score	.88	.66	.02	.86
PSYCHOMOTOR TESTS				
Target Tracking 1 - Mean Log Dist	.97	.68	.23	.74
Target Tracking 1 - Mean Log Dist	.97	.77	.17	.80
Target Shoot - Mean Time to Fire	.91	.48	.06	.85
Target Shoot - Mean Log Dist	.86	.58	.11	.75

reliability are percent correct scores or scores with low split-half reliability.

Overlap with the ASVAB

The squared multiple correlation (SMC) between each test score and the 10 ASVAB subtests is also displayed in Table 2. These SMCs have been adjusted for shrinkage. Only for one test, Number Memory, does the SMC exceed .25. The median SMC across all 19 test scores is .10.

Table 2 also shows the uniqueness for each test score. This value represents an index of the unique (i.e, uncorrelated with the ASVAB) reliable variance of each test score. It is computed by subtracting the SMC with the ASVAB from the split-half reliability. All but two of the uniquenesses in Table 2 exceed .60. This information indicates that these 10 tests have much unique, reliable variance that may contribute to the prediction of job performance.

Overlap among the Computer Tests

Table 3 contains the intercorrelations among the 19 computer test scores. Well over half the intercorrelations between scores on different tests are less than .25, indicating that the various tests are measuring several different abilities.

To determine how we had fared in measuring our target constructs, a principal axis factor analysis was executed. Variables included 17 of the computer test scores (two variables, Perceptual Speed & Accuracy Mean RT and Short Term Memory Mean RT were withheld from the analysis since they correlated .82 and .83 with Perceptual Speed & Accuracy Slope and Short Term

TABLE 3

Intercorrelations among the ASVAB and Pilot Trial Battery (PTB) Tests
Ft. Knox Sample (N=168)

		SRT-RT	CRT-RT	PS&A-PC	PS&A-RT	PS&A-Slp	PS&A-Int	Targ ID-PC	Targ ID-RT	STM-PC	STM-RT	STM-Slp	STM-Int	Can Shoot	No Mem-PC	No Mem-RT	Trk 1-Dist	Trk 2-Dist	TSht-Time	TSht-Dist
PTB - Computerized Cognitive-Perceptual Tests	SRT-RT																			
	CRT-RT	53																		
	PS&A-PC	17	17																	
	PS&A-RT	19	31	50																
	PS&A-Slp	-03	09	52	82															
	PS&A-Int	32	31	-27	-08	-61														
	Targ ID-PC	11	07	40	33	32	-13													
	Targ ID-RT	23	42	20	47	32	13	16												
	STM-PC	-06	04	50	17	26	-23	25	03											
	STM-RT	23	40	26	49	25	23	27	47	08										
	STM-Slp	-06	03	28	29	26	-11	18	13	32	39									
	STM-Int	28	42	10	35	11	31	18	42	-11	83	-19								
	Can Shoot	13	10	00	08	-01	11	-09	25	-02	25	14	18							
	No Mem-PC	-16	-09	29	02	15	-20	14	-12	23	-00	02	-01	-18						
	No Mem-RT	21	24	11	34	21	03	10	27	07	18	15	11	08	-45					
PTB - Computerized Psychomotor Tests	Trk 1-Dist	14	25	-12	08	00	11	-04	42	-29	25	-15	35	27	-14	-00				
	Trk 2-Dist	11	19	-01	11	04	09	02	39	-19	25	-01	27	30	-14	02	81			
	TSht-Time	08	16	16	22	09	12	12	32	09	22	15	15	12	-10	15	23	19		
	TSht-Dist	08	16	-07	03	-00	09	-11	32	-12	27	-16	38	25	-08	02	60	55	-15	

Memory Intercept, respectively), scores from the 10 paper-and-pencil tests described by Toquam et al. (1985), and scores from the 10 ASVAB sub-tests. The sample included only those 168 subjects for whom complete data from all three sets of tests were available. Factor solutions were rotated using the VARIMAX method.

The 7-factor solution was judged the most interpretable. Significant loadings (i.e, greater than .35) for each test score on each factor are shown in Table 4. Based on the factor loadings, we named Factors I-VII general ability, spatial ability, psychomotor ability, general accuracy, basic processing speed, number facility, and a response style factor, respectively. For four of the seven factors (psychomotor ability, general accuracy, basic processing speed, and the response style factor), no paper-and-pencil tests load significantly on these factors. All but one of the tests with significant loadings on the spatial ability factor were paper-and-pencil tests. Both the ASVAB and the computer battery included tests with significant loadings on the other two factors, general ability and number operations; however, the only computer test scores with significant loadings on these factors was Number Memory. Thus, once again, Number Memory appears to be the only computer test that overlaps significantly with the ASVAB.

Some of the factors that include computer tests are moderately similar to the target constructs that we set out to measure with the computer battery. Basic processing speed, for example, contains measures from three target constructs: reaction time, perceptual speed and accuracy, and memory. The number facility factor includes Number Memory test scores, as we had hoped, and also includes the Coding Speed and Number Operations sub-tests from the ASVAB. Finally, the psychomotor ability factor includes

measures of both our target psychomotor ability constructs, multilimb coordination and steadiness/precision.

As Table 4 shows, the time score from Cannon Shoot failed to load significantly on any of the five factors. This indicates that the movement judgment ability tapped by this test differs from the abilities assessed by the other computer tests. This provides indirect evidence that the movement judgment test is measuring a unique perceptual ability, as we had hoped it would.

TABLE 4

Results from a Principal Components Factor Analysis of Scores on the ASVAB, Cognitive Paper-and-Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer Tests[a]
(N = 168)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	h^2
ASVAB GS	75							59
ASVAB AR	75							73
ASVAB WK	77							62
ASVAB PC	62							47
ASVAB NO						84		77
ASVAB CS						62		44
ASVAB AS	62							58
ASVAB MK	77							70
ASVAB MC	63	38	-30					68
ASVAB EI	72							65
Assemb Obj	35	69						66
Obj Rotation		61						49
Shapes		66						51
Mazes		70						67
Path		67	-30					65
Reason 1	37	58						54
Reason 2	37	47						44
Orient 1	37	64						58
Orient 2	40	46			-30			52
Orient 3	60	52						67
SRT-RT					63			44
CRT-RT					61			50
PS&A PC				67	31			70
PS&A Slope				88				81
PS&A Inter				-65	50			74
Target ID PC				40				25
Target ID RT		41	37		30			57
SIM PC				39			34	41
SIM Slope							41	25
SIM Int			38		51			47
Cannon Shoot			32					19
NM PC	53					37		52
NM RT	57					46		54
Tracking 1			86					82
Tracking 2			77					66
Target Shoot TF							42	23
Target Shoot Dist			64					48
Variance Explained	5.69	4.70	2.83	2.37	1.92	1.87	1.17	

[a]Note that the following variables were not included in this factor analysis: APQT, PS&A Reaction Time and Short Term Memory Reaction Time.

(Please also note that decimals have been omitted.)

Discussion

All of the tests except Simple Reaction Time yielded at least one test score with split-half reliability in excess of .80 and test-retest reliability in excess of .55. Thus, we met our first goal, which was to ensure that all the computer tests attained adequate levels of reliability.

Our second goal was to ensure that the new computer tests were not redundant with the ASVAB. SMCs between the 19 test scores and the ASVAB tended to be quite low. Uniquenesses indicated that the computer tests had the potential to contribute a great deal of unique, reliable variance to the prediction of job performance. Thus, we also met our second goal.

Analyses designed to evaluate the intercorrelations among the new tests showed that the various tests generally shared little common variance. Results from a factor analysis indicate that there were at least five (and probably six) different ability factors underlying performance on the 10 tests; these factors are moderately similar to the target constructs we set out to measure. It is important to note here that results from the factor analysis must be considered tentative at best because the sample size includes only 168 subjects. Data obtained from the ongoing concurrent validity study with over 10,000 subjects will provide us with more stable information about our constructs and the relationships among those constructs.

Generally, we felt that the results of the pilot testing indicated that only minor modifications were required in the tests prior to concurrent validation testing. Our observations of subjects during pilot testing suggested a number of changes in the instructions for virtually all of the tests. The split-half reliability data indicated that several of the tests could be shortened without any significant impact on test reliability. Finally, there was some evidence (not discussed in this paper, but noted in McHenry & McGue, 1985) that the two Target Tracking Tests should be made more difficult and that the Target Shoot Test should be made easier. Aside from these, few modifications were made in the computer battery prior to concurrent validation testing. (See Toquam, Dunnette, Corpe, McHenry, Keyes, McGue, Houston, Russell & Hanson, 1985, for more detailed information regarding changes in the computerized perceptual tests.)

Presently, concurrent validation testing is winding down. By the middle of next month, we will have collected predictor and criterion data on almost 10,000 first-term Army enlisted personnel in 19 MOS. It is our hope that at this time next year, we will be able to present some initial validity data for our 10 new computerized perceptual and psychomotor tests.

References

McHenry, J. J., & McGue, M. K. (1985). *Problems, issues, and results in the development of computerized psychomotor measures.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., McHenry, J. J., Keyes, M. A., McGue, M. K., Houston, J. S., Russell, T. L., & Hanson, M. A. (1985). *Development of cognitive/perceptual measures: Supplementing the ASVAB.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., & Houston, J. S. (1985). *Adding to the ASVAB: Cognitive/perceptual measures.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

ADDING TO THE ASVAB: COGNITIVE PAPER-AND-PENCIL MEASURES

Jody L. Toquam, Marvin D. Dunnette, VyVy A. Corpe, and
Janis Houston

Personnel Decisions Research Institute

Introduction

The purpose of this paper is to (1) identify the cognitive/perceptual ability constructs that supplement or provide information about Army applicants' abilities not currently tapped by the Armed Services Vocational Aptitude Battery, or ASVAB; (2) describe the measures developed for paper-and-pencil administration and the cognitive/perceptual constructs they are designed to tap: (3) describe test development issues and the factors used to evaluate the psychometric quality of the new paper-and-pencil measures; and (4) report the relationships between scores on the ASVAB and scores on the new paper-and-pencil tests. Information about the cognitive/perceptual measures designed for computer administration are described in McHenry and Toquam (1985).

Before describing the new tests, we first examine the content of the current military selection and classification battery, the ASVAB, and then provide a brief review of the process involved in identifying the constructs for inclusion in the Pilot Trial Battery. (The Pilot Trial Battery is the term used for the battery of experimental tests administered at Minneapolis MEPS, Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. This battery includes twelve paper-and-pencil measures - ten cognitive and two non-cognitive, and ten computerized measures - seven cognitive/perceptual and three psychomotor.)

The current military selection and classification battery, the ASVAB, contains ten subtests. Scores on four of these are used to calculate the Armed Forces Qualification Test (AFQT) score which is used to determine qualification for entrance into the Army. Scores on the ten subtests are used in different combinations to determine applicants' qualifications for different military occupational specialties (MOS). Results from a factor analysis of ASVAB scores indicate that the battery assessed verbal ability, speeded performance, quantitative ability, and technical knowledge (Kass, Mitchell, Grafton & Wing, 1982).

Peterson (1985) describes the activities involved in identifying ability constructs that supplement information obtained from the ASVAB. Those activities included a review of the literature which was used to impose structure on the domain (i.e., establish a cognitive/perceptual abilities taxonomy) and then to summarize validity data for the different types of ability constructs. This information was input to the expert judgment task. All of this information was used to identify cognitive/perceptual ability constructs that tap abilities relatively independent of those measured by the ASVAB and that may be used to improve the Army's selection and classification decisions process.

Cognitive/perceptual ability constructs selected for inclusion in the Pilot Trial Battery and their designated priorities (in parentheses) are: (1) Spatial Visualization - Rotation and Scanning;

(2) Spatial Visualization - Field Independence; (3) Spatial Orientation; (4) Induction - Figural Reasoning; (5) Reaction Time - Processing Efficiency; (6) Memory - Number Operations; (7) Memory - Short Term Memory; (8) Perceptual Speed and Accuracy.

Determining the Method of Administration

In this section, we review the factors that influenced our decision to measure a particular construct via paper-and-pencil or via computer. The first factor concerns the construct definition and the dependent measures suggested by that definition. For example, definition of the construct, processing efficiency, indicates that the dependent measure involves the time required to respond to simple stimuli. Such information can only be obtained on a computer because a precise measure of reaction time is required. Hence, those constructs that involve a reaction time component, such as Processing Efficiency, Perceptual Speed and Accuracy, and Memory were slated for computer administration. McHenry and Toquam (1985) provide a detailed description of measures developed for computer administration.

The second factor involves the cost related to adapting items to the computer. For example, test items for such constructs as spatial visualization and figural reasoning involve detailed figures and objects. To adapt these items to the computer would require high resolution graphics. The cost for hardware capable of supporting such graphics at the time was prohibitive. Thus, we determined that measures of spatial visualization, spatial orientation, and induction would be assessed via paper-and-pencil. We focus on the development activities and pilot-test results of the new paper-and-pencil measures in the remainder of this paper.

Paper-and-Pencil Measures: Construct and Test Descriptions

In this section, we provide definitions of the constructs, describe criterion job performance areas or tasks that we expect measures of the constructs to predict and finally identify the tests designed to measure each construct. Detailed descriptions of the individual tests are available from the authors.

Spatial Visualization--Rotation

This involves the ability to mentally restructure or manipulate parts of a two- or three-dimensional figure. It serves as a potentially effective predictor of success in MOS that involve mechanical operations, construction and drawing or using maps. Two tests developed to measure this construct include Assembling Objects and Object Rotation.

Spatial Visualization--Scanning

This includes the ability to visually survey a complex field and to find a pathway through it. According to our expert judges, measures of this construct are potentially effective as predictors of success for Army MOS involving electrical or electronics operations, using maps in the field, and controlling air traffic. The two measures designed to assess this construct in the Path Test and the Maze Test.

Spatial Visualization--Field Independence

This includes the ability to find a simple form when it is hidden in a complex pattern. A measure of this construct is expected to predict success in MOS that involve detecting and identifying targets, using maps in the field, planning placement of tactical positions, air traffic control and troubleshooting operating systems. The Shapes Test was developed to measure this construct.

Spatial Orientation

This involves the ability to maintain one's bearing with respect to points on a compass and to maintain appreciation of one's location relative to landmarks in the environment. From job observations conducted in the field, we expect measures of this construct to predict success in combat MOS that involve maintaining directional orientation using features of landmarks in the environment. Three tests involving different orientation tasks were developed to assess this construct, Orientation 1, Orientation 2, and Orientation 3.

Induction - Figural Reasoning

This includes the ability to generate hypotheses about principles governing relationships among several objects. According to the panel of experts, measures of this construct are effective predictors of success in MOS involving troubleshooting, inspecting, and repairing electrical, mechanical, or electronic systems, analyzing data, controlling air traffic, and detecting and identifying targets. We developed two tests involving different tasks to assess abilities in this construct area. These were titled Reasoning 1 and Reasoning 2.

Test Development Issues

Two issues impacted on our approach for developing the new paper-and-pencil measures. These include the target population completing the new tests for selection and classification purposes and the power versus speed components of each new test. We discuss each in turn below. The population completing these tests is the same population that completes the ASVAB to qualify for entrance into to the Army. This is, very generally speaking, a population composed of predominantly recent high school graduates, not entering college, from all geographic sections of the United States. For our purposes the target population was, practically speaking, inaccessible during the test development phase. We were constrained to using enlisted soldiers to try out the newly developed tests. The development group, enlisted soldiers, of course, represents a restricted sample because they have passed enlistment standards and often have completed basic and advanced individual training.

Differences between the target population and the sample available to us, lead to two major implications that served as general guidelines for test development and pilot testing activities. First, the target population includes a broad range of abilities, therefore we attempted to develop test with a broad range of item difficulties. And second, the the test development group, first-term enlistees, would be of generally higher in ability than the target population. Therefore, the overall difficulty level of the test should be somewhat higher (i.e.,the test should be somewhat easier) than what it would have been if we had access to an unrestricted sample of the target population.

Another decision to be made about each test was its placement of the power vs. speed continuum. Most psychometricians would agree that a "pure" power test is a test administered such that all persons taking the test are allowed enough time to attempt all items on the test, and that a "pure" speeded test is a test administered such that no one or very few taking the test has enough time to attempt all items. In practice, there appears to be a power/speed continuum, most tests fall somewhere between the two extremes on this continuum.

During the preliminary test development stage, we categorized each test as a power test, speeded test, or combination of the two using our construct definitions. For example, using our definition of Induction, we designed the test items to represent a very wide range of difficulty levels and established a generous time limit such that most subjects would have time to complete all items. Thus, measures of induction were designed to fall on the power end of the continuum. Our plan for measures tapping Spatial Visualization -Rotation and Scanning differed from this in that all items were constructed to be moderately easy but more restrictive time limits were imposed. Thus, these measures were intended to fall toward the speeded end of the continuum.

For the remaining constructs, Spatial Visualization-Field Independence and Spatial Orientation, we designed the measures using the construct definitions to determine the range of item difficulties and to establish time limits. Following each pilot-test we examined completion rates and item difficulty levels to assess how closely performance on each new test matched the corresponding construct definition with regards to speed and power components.

Evaluating the Paper-and-Pencil Tests

Four pilot test or tryout sessions were conducted at Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. In the first pilot-test at Fort Carson, about 38 soldiers completed each paper-and-pencil test. The number at Fort Campbell was 57 and at Fort Lewis it was 118. At Fort Knox the numbers were 290 for time one and 97 to 126 for time two. Factors used to evaluate each test at one or more of these pilot-test sessions include the following: construct validity, test item characteristics, and test reliability. Below we present some general findings for all paper-and-pencil tests.

One goal of the the pilot-test sessions was to verify the construct validity of the new measures. Therefore, we identified published tests that measure constructs similar to our construct definitions. These published measures were included in the first three pilot-tests. It is important to note that, in general, most published tests or marker tests differed from the new tests in item difficulty levels and in the specific task required. Therefore, we did not expect a one-to-one correspondence between the new test and its published marker test.

Very few of the newly developed tests correlated above .65 with the designated marker; most correlations between new measures and marker tests fell between .45 and .60. These values were as expected given the differences in task requirements and in item difficulty levels between the new and marker tests. Basically this information suggested to us that although the tests did not duplicate their respective marker tests, they captured the essence of the target con-

struct.

Another goal of the pilot-test sessions was to assess the psychometric characteristics of each new test. Following the pilot-test sessions, then, we computed item difficulty levels and item-total correlations for each test. These data were used to modify test items and to adjust time limits.

Results from the first pilot test indicated that all tests required some modification. For example, completion rates, item difficulty levels and raw total test scores suggested that some of the new measures may suffer from ceiling effects. Thus, for Assembling Objects, Object Rotation, Path Test, and Orientation 1, we constructed new items and adjusted time limits accordingly to obtain the desired difficulty level. For the Shapes Test and Maze Test, we modified test items to increase difficulty levels and to reduce the possibility of ceiling effects.

The reverse situation appeared on one of the orientation tests, Orientation 1. That is, item difficulty levels were low or the test was more difficult than desired. We modified this test by adding four "easy" items and by expanding the time limit.

For the remaining measures, Orientation 3, Reasoning 1, and Reasoning 2 very few changes were required. For example, item analysis data revealed that for some of the items, item-total correlations were higher for a distractor than for the correct response. These items were either modified or replaced.

Subsequent pilot tests indicated that the tests, in general, required only minor modifications.

Finally, we investigated the reliability or internal consistency and the stability of each new measure. To compute internal consistency estimates we used a split half procedure. This included administering each test as two separately timed halves and computing the correlation between part one and part two for each test. The Spearman-Brown correction procedure was then used to estimate the reliability for the test as a whole. We estimated the stability of each test by collecting test-retest data on a sample of about 100 soldiers at Fort Knox. A period of two weeks separated the two test sessions.

Internal consistency and test-retest estimates for each test appear in Table 1. Results from the Fort Lewis pilot-test indicate that the split half internal consistency estimates range from the high 70's to the low 90's for all tests with the exception of Reasoning 2. Test-retest estimates are lower than the internal consistency estimates but are at acceptable levels ranging from .57 to .84. The Reasoning 2 test once again yields the lowest value of all.

Note that in Table 1, we have also included internal consistency estimates for the Fort Knox sample computed using the Hoyt formula. and may represent overestimates for some of the more highly speeded tests. With the exception of Reasoning 2, these values range from the low 80's to high 90's.

TABLE 1

Reliability and Uniqueness Estimates for the Ten Paper-and-Pencil Tests Included in the Pilot Trial Battery

			Reliability			Uniqueness	
			Ft. Lewis	Ft. Knox			
Test	No. Items	Time Allotted (in minutes)	r_{xx} Split Half N = 118	Alpha N = 790	r_{xx} Test-retest (N = 97 to 126)	ASVAB R^2 Using Split Half	ASVAB U^2 Using Split Half
Assembling Objects	40	16	.79	.92	.74	.40	.39
Object Rotation	90	7.5	.86	.97	.75	.19	.67
Mazes	24	5.5	.78	.89	.71	.25	.53
Path	48	8	.82	.92	.64	.29	.53
Shapes	54	16	.82	.92	.70	.19	.63
Reasoning 1	30	12	.78	.83	.64	.29	.49
Reasoning 2	32	10	.63	.65	.57	.26	.37
Orientation 1	150	10	.92	.98	.67	.36	.56
Orientation 2	24	10	.89	.88	.80	.30	.59
Orientation 3	20	12	.88	.90	.84	.54	.34

Overlap Between the New Measures and the ASVAB

As we above, the major focus of this research involves identifying and developing measures of constructs not currently assessed in the ASVAB. One way to estimate the amount of overlap between each new measure and the measures contained in the ASVAB is to conduct uniqueness analyses. This procedure involves computing the squared multiple correlation between each new test and the ten ASVAB subtests. The resulting value is then subtracted from the reliable variance in that new measure (in this case we used the reliability estimate computed using the split half procedure). This value represents an index of the unique variance or variance that is uncorrelated with scores obtained on the ASVAB. Results from this analysis are reported in the final two columns in Table 1.

Across the ten new tests, the squared multiple correlations range from .54 to .19. It is clear that some of these tests are measuring abilities tapped by ASVAB subtests. On the other hand, the uniqueness estimates which range from .67 to .34, indicate that the new tests tap abilities independent from those assessed by the ASVAB subtests.

In sum, results from the uniqueness analysis are essentially what we would expect in assessing the amount of overlap between groups of tests that measure cognitive/perceptual abilities. The data are encouraging because they indicate that we are measuring ability constructs not currently assessed by the ASVAB.

REFERENCES

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1982). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10: 1981 Army applicant sample, (Technical Report 581). Alexandria, VA: U.S. Army Research Institute.

McHenry, J. J., & Toquam, J. L. (1985). Computerized assessment of perceptual and psychomotor abilities. Paper presented at the Military Testing Association Conference, 24 October, San Diego, CA.

Peterson, N. G. (1985). Mapping predictors to criterion space: Overview. Paper presented at the Military Testing Association Conference, 24 October, San Diego.

Measuring Personal Attributes: Temperament, Biodata, and Interests

Leaetta M. Hough, Matt K. McGue, John D. Kamp,
Janis S. Houston, and Bruce N. Barge

Personnel Decisions Research Institute

Overview. I'm going to describe the development and evaluation of temperament, biographical, and interest measures - what we call non-cognitive measures - included in the Project A predictor battery. Non-cognitive measures were included in the predictor battery because of their potential for predicting important on-the-job criteria, criteria such as Effort, Initiative, Following Regulations and Orders, Adjustment, Leadership, and Self-Control.

The information I will present today suggests: 1) that non-cognitive predictors are likely to predict such criteria; in fact, more likely to predict such criteria than are other types of predictors; 2) that non-cognitive measures contribute unique variance to the predictor battery and are, therefore, likely to contribute incremental validity; 3) that the non-cognitive measures we developed have good psychometric characteristics, they are internally consistent and show high test-retest reliability; and 4) that faking on personality inventories is not the problem it is often assumed to be. Our overall strategy was: to review the literature on temperament, biodata, and interest to identify constructs that were likely to be criterion valid; to obtain expert judgments about expected true validity of those constructs; to develop measures of those constructs; to remove or revise sensitive or objectionable items; and to evaluate and revise measures based on their internal consistency, overlap with other predictors, and their stability across time and different motivational conditions.

Literature Review Results. Our review and summary of the literature indicated that the validity of interest measures for important Army criteria were in the high .20s. The validities of biographical inventories for such criteria were in the .20s and .30s. These results were not too different from previous literature reviews. Our conclusions for the personality literature, however, differ from some of the other reviews, and I'd like to describe these results more thoroughly.

The criterion-related validities reported in the literature for temperament constructs are shown in Table 1. As you can see, the adjustment criterion, which includes such things as unfavorable discharge and drug abuse, is predicted very well by temperament measures. The predictor constructs Achievement and Locus of Control also predict Educational, Training, and Job Proficiency criteria. These results differ from those reported by Guion and Gottier in their 1965 Personnel Psychology article. Our results are, however, similar to those reported by Ghiselli in his 1973 Personnel Psychology article. We believe the results are explained by the approach we used.

Our approach was to develop a predictor taxonomy and to classify temperament scales into the taxon or construct with which they were most similar. We accomplished this classification by searching the literature

Table 1

Summary[a] of Criterion-Related Validities of Temperament Constructs

Temperament Construct	Type of Criterion: Educational	Training	Job Proficiency	Job Involvement	Adjustment
Potency (Surgency)	.06 (42)[b]	.13 (36)	.07 (65)	.04 (13)	-.17 (31)
Adjustment	.14 (43)	.19 (28)	.11 (65)	.17 (16)	**-.33** (52)
Agreeableness (Likeability)	.03 (9)	.08 (5)	.03 (22)	-.02 (5)	-.03 (5)
Dependability	.13 (24)	.12 (20)	.11 (49)	.14 (15)	**-.43** (40)
Intellectance (Culture)	.17 (6)	.19 (5)	.01 (16)	-.09 (9)	.18 (3)
Affiliation	-.03 (5)	--- ---	-.02 (6)	.09 (4)	-.07 (4)
Achievement	**.30** (8)	**.33** (4)	**.24** (4)	--- ---	**-.33** (5)
Masculinity	-.16 (8)	.09 (3)	.10 (10)	.03 (4)	-.13 (11)
Locus of Control	**.32** (1)	**.29** (2)	**.25** (7)	--- ---	--- ---
Unclassified Military Scales	--- ---	.18 (8)	.18 (25)	--- ---	**-.22** (20)

[a] Medians are reported as the summary index.
[b] The number in parentheses is the number of correlations on which the median is based.
NOTE: Median correlations greater than .20 are indicated by a box.

for reported correlations between temperament scales and then using these correlations to categorize the temperament scales into the five factors identified by Tupes and Christal (1961) in their peer rating research. We then added four constructs to the taxonomy to increase the homogeneity of the constructs. We also used a taxonomic system for the criteria. These consisted of Educational, Training, Job Proficiency, and Adjustment criteria.

We then summarized the criterion-related validities reported in the literature according to our predictor and criterion taxonomies. Guion and Gottier did not summarize the literature according to constructs; Ghiselli, however, reported results only for studies for which he felt the predictor was conceptually appropriate for the criterion. Our literature review, which summarized the reported validities according to a data-based classification of scales into constructs, supports Ghiselli's results and conclusions. We believe the construct approach highlighted the predictor-criterion relationships by reducing the "noise," if you will, and that the Guion and Gottier approach masked such relationships.

Expert Judgments of True Validity. Using the construct approach, we identified the temperament constructs that were likely to yield good criterion-related validities. We then asked experts to estimate the expected true criterion-related validities of predictor constructs for important Army criteria. These estimated validities also indicated that the non-cognitive predictors were likely to predict Army criteria - criteria such as Initiative/Effort, Following Regulations and Orders, Leading and Supporting, Self-Control, and others in the .20s, .30s, and even .40s. I might add that the cognitive and psychomotor measures were not expected to predict these criteria nearly as well.

Development of Construct Measures. Using the results of the literature review and expert judgments, we identified "good bets" for predicting important Army criteria. We developed scales to measure these constructs.

We wrote temperament and biodata items for the ABLE, which stands for Assessment of Background and Life Experiences, and we wrote interest and biodata items for the AVOICE, which stands for Army Vocational Interest Career Examination. We also developed four "response validity scales" which we called Social Desirability, Poor Impression, Self-Knowledge, and Non-Random Responses and included the items in these four response validity scales in the ABLE.

We next examined the ABLE and AVOICE items for sensitivity, or the extent to which people might object to the content of the questions. The Army and their scientific advisors also reviewed the items for sensitive content. We revised or removed the objectionable items and administered the ABLE and AVOICE to soldiers at Ft. Lewis, Ft. Campbell, and Ft. Knox. After each administration we examined the psychometric characteristics of the items and scales and revised them for each subsequent administration.

The last administration was at Ft. Knox where about 275 soldiers completed the ABLE and AVOICE. We evaluated the scales for internal consistency, test-retest reliability, and their unique contribution to the predictor battery. For the ABLE scales, the median internal consistency was .84, with a range of .70 to .87. For the AVOICE, the median was .86, with a range of .68 to .96. About 125 soldiers returned two weeks later to complete the ABLE and AVOICE a second time. The median test-retest coefficient for the ABLE was .79, with a range of .68 to .83. For the AVOICE, the median test-retest was .76, with a range of .56 to .86. Uniqueness analyses we conducted show that both the ABLE and AVOICE share very little variance with the ASVAB or with the cognitive and psychomotor tests included in the predictor battery. In short, the psychomotor characteristics of both the ABLE and AVOICE are very good; they are internally consistent, stable over time, and likely to contribute incremental validity to the predictor battery.

Faking Study. The next issue we addressed was faking. The concern was that self-report measures are susceptible to intentional distortion. We, therefore, conducted a faking study, the purpose of which was 1) to determine the extent to which soldiers can distort their responses to temperament and interest inventories when instructed to do so; 2) to determine the extent to which the ABLE response validity scales detect intentional distortion; 3) to determine the extent ABLE response validity scales can be used to adjust or correct scores for intentional distortion; and 4) to determine the extent to which distortion is a problem in an applicant setting.

We gathered data from 125 Army applicants - people who wanted to be accepted into the Army and would have a motive for distorting their responses; we used the Ft. Knox data as an honest comparison sample; and we conducted an experiment in which soldiers were instructed to respond honestly or to distort their responses in a specified way.

The participants in the experimental group were 245 enlisted soldiers at Ft. Bragg. We created four faking conditions: fake good on the ABLE, fake bad on the ABLE, fake interest in combat activities on the AVOICE, and fake interest in non-combat activities on the AVOICE. We also created two honest conditions: honest on the ABLE, and honest on the AVOICE.

The design was a repeated measures with faking and honest conditions counter-balanced. Thus, approximately half the experimental group, or 124 soldiers, completed the inventories honestly in the morning and faked in the afternoon, while the other half (121 soldiers) completed the inventories honestly in the afternoon and faked in the morning. In summary then, we had a 2 x 2 x 2 fixed-factor, completely crossed experimental design.

We performed a multivariate analysis of variance on the ABLE and AVOICE scales separately. All the relevant fake x set interactions for the ABLE were significant at the .01 level, indicating that soldiers can distort their responses. The fake x set x order interactions, significant at the .05 level, indicate that the order in which the conditions occurred has a significant effect on scores. We performed a multivariate analysis of variance on the AVOICE scales and found similar results; people can distort their responses to an interest inventory.

Another research question was the extent to which the response validity scales detected intentional distortion. The results indicate that the Social Desirability scale detects faking good; the effect size of the difference between the means for the honest and fake good conditions is 1.02, or one standard deviation. The Poor Impression scale detects faking bad; the effect size of the difference between the means for the honest and fake bad conditions is 2.67, or just over two and one-half standard deviations.

We next examined the extent to which we could use the response validity scales, Social Desirability and Poor Impression, to adjust ABLE content scales and AVOICE occupational scales for faking. We regressed out Social Desirability from the fake good condition and Poor Impression from the fake bad condition. Table 2 shows the median effect sizes between the honest and faking conditions for the ABLE and AVOICE scales before and after regressing out Social Desirability and Poor Impression. The median difference in ABLE scores between the honest and fake good condition before regressing out Social Desirability is .49 or half a standard deviation. That is, ABLE scale scores differ by about half a standard deviation in the fake good condition as compared to the honest condition. After regressing out Social Desirability from the fake good condition, the ABLE content scales are only .14, or just over 1/10 of a standard deviation, different from the honest condition.

The median difference in ABLE scores between honest and fake bad before regressing out Poor Impression for is 2.10. That is, ABLE content scale scores in the fake bad condition differ by approximately two standard deviations from ABLE content scales in the honest condition. However, after regressing out Poor Impression from the scales, the difference is less than half a standard deviation. Clearly, the response validity scales Social Desirability and Poor Impression can be used to adjust scale scores for the ABLE for intentional distortion. We do not know, however, whether the adjustment formula will cross-validate and be as effective in another data set. Nor do we know whether adjusting the scale scores improves the criterion-related validity of the scales. It may be that the unadjusted scale scores are more criterion-valid than adjusted scores.

We performed the same computations for the AVOICE occupational scales and

Table 2

Effects of Regressing Out Response Validity Scales
(Social Desirability and Poor Impression)
in Faking Conditions for ABLE and AVOICE

	Honest vs Fake Good/Combat Effect Size		Honest vs Fake Bad/Non-Combat Effect Size	
	Before Adjustment	After Adjustment	Before Adjustment	After Adjustment
ABLE Content Scales	.49	.14	2.10	.45
AVOICE Combat Scales	.43	.33	.97	.86
AVOICE Combat-Support Scales	.55	.39	.49	.34

Median values are reported.

found that the results are not nearly as impressive. The bottom two rows show the median effect size of the differences between the honest and faking conditions before and after regressing out the appropriate response validity scale for the AVOICE.

These data demonstrate that: 1) people can distort their responses to temperament and interest scales, 2) response validity scales detect such distortion, and 3) the response validity scales can be used to adjust temperament scale scores for distortion. However, the question remains: To what extent do applicants distort their responses? To answer this question we compared scale scores from the Ft. Bragg experimental honest condition and the Ft. Knox honest condition with the scale scores of approximately 120 Army applicants. These comparisons suggest that applicants do not appear to distort their responses. As shown in Table 3, the applicant means on the temperament scales (ABLE content scales) are lower than one or both of the honest means nine out of eleven times. The results for the AVOICE are similar. In short, applicants do not tent to distort their responses.

Summary. To briefly summarize our approach and results: we identified constructs and developed measures of constructs that had demonstrated criterion-related validity in the past and were judged by expects as likely to be criterion-valid for important Army criteria. The measures we developed contributed unique variance to the predictor battery, were internally consistent or homogeneous, and yielded reliable and stable scale scores across time and motivational conditions.

Our next step is to criterion-validate these measures with Army criteria. Data gathering for that is currently underway.

Table 3

Comparison of Ft. Bragg Honest*, Ft. Knox, and MEPS (Applicants) ABLE Scales

ABLE Scale	Ft. Bragg (Honest)* N	Ft. Bragg (Honest)* Mean	MEPS (Applicants) N	MEPS (Applicants) Mean	Ft. Knox N	Ft. Knox Mean	Total S.D.
Response Validity Scales							
Social Desirability (Unlikely Virtues)	116	15.91	121	16.63	276	16.60	3.21
Self-Knowledge	116	29.54	121	28.03	276	29.64	3.63
Non-Random Response	116	7.58	121	7.79	276	7.75	.64
Poor Impression	116	1.50	121	1.05	276	1.54	1.84
Content Scales							
Emotional Stability	112	66.22	118	66.03	272	65.05	7.86
Self-Esteem	112	34.77	118	34.04	272	35.12	5.00
Cooperativeness	112	53.33	118	54.60	272	54.19	6.05
Conscientiousness	112	46.37	118	46.49	272	48.97	5.86
Non-Delinquency	112	53.24	118	54.36	272	55.49	6.91
Traditional Values	112	36.67	118	36.97	272	37.28	4.50
Work Orientation	112	59.71	118	58.37	272	61.40	7.73
Internal Control	112	49.48	118	51.90	272	50.37	6.13
Energy Level	112	57.56	118	56.67	272	57.19	6.95
Dominance (Leadership)	112	35.54	118	32.84	272	35.41	6.05
Physical Condition	112	32.96	118	28.27	272	31.08	7.49

*Scores are based on persons who responded to the honest condition first.

References

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Guion, R. M., & Gottier. (1965). Validity of personality measures in personnel section. Personnel Psychology, 18, 135-164.

Tupes, E. C., & Christal, R. E. (May, 1961). Recurrent personality factors based on trait ratings (ASD-TR-61-97). Lackland Air Force Base, TX: Aeronautical Systems Division, Personnel Laboratory.

SYMPOSIUM OVERVIEW

THE TRAINING AND SELECTION OF ARMY MANAGERS: QUANTITATIVE/QUALITATIVE APPROACHES

Gerald P. Fisher, Human Resources Research Organization (HumRRO)

Richard Lilienthal, Army Civilian Personnel Center (CIVPERCEN)

The U.S. Army's civilian workforce currently exceeds 450,000 employees. As this widely dispersed workforce becomes increasingly dependent on highly skilled employees capable of keeping pace with rapidly advancing technologies, major personnel management challenges are presented. Among those challenges are the needs of a personnel system to select, train, assign, promote and retrain employees and managers within the Army. The presentations this morning include the following: Selecting and Training Logistics Managers; Designing an Executive Development Program; and Managerial Competencies for LOGAMP.

This Military Testing Association symposium concerns three projects related to training and selecting Army civilian professionals, managers and/or executives. In order to promote and develop a diverse and talented workforce, emphasis needs to be placed on recruiting, selecting, promoting and training generalists who can meet the challenge of leadership and change. Each of the three presentations focuses on some aspect of how training requirements can be established and sequenced to effectively train specialist supervisors in the direction of generalist managers. Through promotion, cross-training, management development programs, and other methods, it is presumed that the goal of selecting and producing a well trained generalist can be accomplished.

At the same time, the technical sophistication of the equipment and technology that the new Army must deal with today and tomorrow demands that civilian professionals and managers have state-of-the-art training, the capability of handling increasingly complex concepts and the skills of a specialist. The generalist/specialist dichotomy runs through each of our presentations. Which skills are essential and what knowledge(s) must be possessed depend on using reliable and defensible data gathering and decision making methods. Thus each of these three presentations relies on both in-person interviews as well as questionnaires and formal data gathering instruments. While sample size varied from 10,000 incumbents in the logistics study to as few as 10 executives in the finance center effort, each of the studies does rely on more than one data gathering method.

All presenters this morning are outside the field for which they are recommending training or selection criteria. Each of the papers will hopefully shed light on two related issues. The first is the question of the qualitative role of the analyst at key stages of each of the studies. Each of us in the symposium recognizes that when we ask a question of a group of job incumbents that we in some way influence the response that is received.

When we analyze a data set we also bring to bear our experiences and perspectives. One of the issues in this panel is the role of the analyst and how his or her experience and perspectives influence eventual policy recommendations and training or selection decisions. A second issue of concern in these presentations is that of sample size and the value and limits in using either or both large and small sample data gathering methods. Each of the speakers will discuss this variable in our presentations.

One of the key issues addressed in the logistics study (Lilienthal, Fisher, Hough) is whether a single job analysis can be used effectively for multiple purposes. In our paper we describe a job analysis project in which both training and selection uses evolve from the same task data base. Both management and employees complain about the presence of separate job analyses by hiring, training, and classification personnel. A single procedure that would serve all of management's job analytical needs has been a goal of job analysts for years. This project has not discovered that procedure. It does, however, employ a true multi-purpose procedure in that the products will be used in both selection and training systems.

The three major uses of job analysis, i.e., selection, classification, and training, have different goals and needs. For selection purposes, the analyst typically looks for similarities among jobs. This is because the more that jobs are found to have in common, the more they can share common selection procedures. General aptitude batteries such as DoD's Armed Forces Vocational Aptitude Battery (ASVAB) and Office of Personnel Management's former Professional and Administrative Career Examination (PACE) could not exist without commonality across a number of jobs. Tasks need not be written at a detailed level of specificity if the resulting knowledges, skills, and abilities (KSAs) are general dimensions such as deductive reasoning, inductive reasoning, spatial, numerical, oral, and writing abilities. At the extreme of this viewpoint are supporters of the validity generalization concept who suggest that a selection-oriented job analysis can be quite simple, needing only to demonstrate that the job under study belongs to a general class of jobs for which selection validation data exist.

For classification and pay purposes, the analyst typically looks for differences among jobs. This is because the analyst must develop or support a system which assumes differences among jobs (e.g., the Federal civilian classification system which assumes that there are hundreds of series, each having multiple pay grades). To obtain these differences, the analyst writes task statements that are specific to series and pay grades.

For training purposes, the analyst needs more detailed tasks and KSAs than for selection purposes. One reason is that the task of the trainer is to bring novices to the full performance level. While the selection specialist can concentrate on only the tasks and KSAs that distinguish superior job performance, the training specialist must be concerned with teaching everything necessary to perform the job. Another reason is that learning objectives are more specific than dimensions such as oral ability or writing ability.

This project uses the same task inventory for selection and training purposes. The dual purpose does not become apparent until the SME panels. The level specificity of the task statements is that required for training.

Although that degree of detail is not required for selection purposes, there is no problem caused by having it. Job classification is not a purpose of this project. If it were, many more tasks delineating level of supervision and responsibility would be added. For selection and training purposes such tasks are not needed because they do not add anything useful to the analysis. For example, the same KSAs are needed to write a report as to review and approve it.

The two purposes manifest themselves in the SME panels, not just in the rating scales used, but in the very nature of the KSAs. There are typically more KSAs identified for training, i.e., they are written at a more detailed level of specificity. Some KSAs appear as products for both purposes but in different forms. For example, in the Army Civilian Personnel Administration career program, Army pay policy might be identified for training purposes while Federal pay policy would be identified for selection purposes. Although we would teach Army policy we would not want to use an Army-specific KSA in the selection system. Agency-specific selection KSAs tend to restrict competition and result in organizational "in breeding." Other KSAs appear as products for only one purposes in any form. For example, we offer courses on stress management but we do not rate and rank applicants on stress management. Because of regulations and litigation, our criteria for KSAs are more strict for selection than training purposes.

In summary, each of these projects found that the former methods of a few decision makers and staff members getting together to decide the content of promotion elements or training requirements has had to be replaced by a more rigorous and legally defensible multi-method approach. We look forward to sharing details of these projects with the readers of the papers and with the audience present at MTA.

SELECTING AND TRAINING LOGISTICS PROFESSIONALS AND MANAGERS: QUALITATIVE/QUANTITATIVE APPROACHES

Gerald Fisher
Human Resources Research Organization

Richard Lilienthal
Army Civilian Personnel Center

Leaetta Hough
Personnel Decisions Research Institute

This paper focuses on methodological issues involved in one of the three studies in this symposium--a Comprehensive Occupational and Data Analysis Programs (CODAP) job analysis producing selection criteria and training requirements for GS/GM 11-15 Army civilians in three career programs--Supply Management, Materiel Maintenance Management, and Transportation Management. To understand the need for this job analysis, it is necessary to outline several recently developed Army managerial training, career development and selection programs.

The Army is in the process of developing two systems for selecting and training civilian managers. The Army Civilian Career Evaluation System (or ACCES) is the result of a joint effort between the U.S. Army Civilian Personnel Center (CIVPERCEN) and the U.S. Office of Personnel Management to improve the current civilian promotion and referral system, known as the SKAP system. ACCES requires a rigorous task analysis effort so that the job tasks required can be quantitatively specified. Once the task clusters are identified, appropriate knowledges, skills, and abilities (KSAs) can be specified for promotion and referral purposes throughout the Army. ACCES will be the Army's future centralized evaluation and referral system and has thus far been implemented in two civilian career programs--Manpower and Force Management and Civilian Personnel Administration. This particular study looks at the problems of developing selection criteria and training requirements in three other career programs within a single effort.

The second purpose for which this job analysis was conducted is the newly developed Army Civilian Training, Education and Development System (ACTEDS). ACTEDS is targeted toward improving the development of the Army's civilian work force through systematic technical, professional, and managerial training and development. The Army recognizes that ACTEDS is needed because:

> As presently designed, the civilian training and development system does not fully support the progressive development of Army's future top civilian managers. Contrary to the desired orderly, systematic approach to technical, professional, and managerial skills training, civilian employees typically participate in programs on a self-initiated rather than management planned basis. In most

instances, the training and assignments they receive are not sequentially interrelated to contribute to progressively increasing and strengthening the experience and knowledge base over their entire career.

One of the first civilian career programs in which ACTEDS is being implemented is the Logistics and Acquisition Management Program (LOGAMP) consisting of all GS/GM 11-15 managers in six selected Army career programs: Contracting and Acquisition; Quality and Reliability Assurance Engineers and Scientists; Material Maintenance Management; Supply Management; and Transportation Management. (The last three career programs noted are being analyzed in this study.) The method for developing ACTEDS training requirements, as well as selection criteria for promotion within the ACCES program, is through a CODAP-based job task analysis inventory followed by subject matter expert (SME) workshops wherein KSAs are developed both for training requirements and for setting selection criteria.

Method

Since September 1984, the Personnel Decisions Research Institute (PDRI), joined by the Human Resources Research Organization (HumRRO), has been conducting a CODAP-based job analysis for the Army Civilian Personnel Center (CIVPERCEN). Individual job task and KSA lists for the 20 job series within the three career programs were initially developed. The lists were based on a review of 2,000 position descriptions (out of 10,000 job incumbents). Using the review of current classification and qualification standards for each series as well as the initial inventories as a starting point, a sample of nearly 400 incumbents was interviewed in small group meetings to add, modify, or eliminate task statements. The analysts merged the individual inventories, resulting in a single job description inventory of more than 300 task statements covering all three career programs. The single LOGAMP task inventory was then distributed to all 10,000-plus job incumbents at Army installations throughout the world. Following receipt of completed inventories, CODAP analyses are to be conducted and pertinent task and duty clusters will be developed and validated by SMEs.

Development of the CODAP Inventory

CODAP was originally developed by the Air Force for occupational analysis purposes in the mid to late 1950s. CODAP includes a set of interrelated statistical programs used for the purpose of job analysis.

The primary focus of CODAP is the analysis of jobs, jobs being defined as a set of tasks selected from a standard inventory. The selection of tasks is performed by the job incumbent to reflect those tasks actually performed by him or her. In addition to identifying the tasks, the job incumbents are asked to categorize the relative time spent performing each task. When these "time spent" categories are assigned numerical weights and converted to percentages, individual job descriptions are produced. In this inventory we asked incumbents to rank both relative time spent (to other tasks) and importance to job on each task performed.

In our current CIVPERCEN project, we plan to analyze the data on the basis of empirical task/duty clusters, rather than on previously developed job series within the various career programs that are being reviewed. Furthermore, we have attempted to go beyond the definition of jobs defined merely by a time spent analysis and have included an importance to the job scale with each task statement. Additionally, because the types of equipment and commodities are thought by incumbents to be such a critical variable (in many cases) for selection and training of high level professionals and managers, we have added a separate section to the job description inventory to identify the equipment/commodites that differentiate various jobs at various levels of the organization. This additional section also allows us to greatly shorten the length of the inventory. For instance, rather than a separate task statement such as "develops long range budget for fixed wing aircraft," as well as a similar statement for each type of equipment covered in the career program(s), we merely listed each commodity and equipment type once and shortened the task statement to read "develops long range budget." As can be envisioned, this shortens the inventory from several thousand items to several hundred (349 task statements along with 88 commodities and equipment categories.)

Once the statistical clusters of tasks and duties are established through the various CODAP computer programs, selected SMEs will review the data and establish the task/duty clusters. Following the specification of task and duty clusters, SME panels will be formed for each job cluster to develop KSAs needed for training requirements and selection criteria.

DISCUSSION

The methodology employed in our study is a multi-method multi-purpose job analysis. The two methods are 1) a task inventory with CODAP analysis and 2) panels of subject-matter-experts (SME). The two purposes are to identify jobs and KSAs for both promotion and training. The task inventory is used to identify and define jobs; the SME panels are used to identify the KSAs for promotion and training.

Multiple Method

The importance of this project required the best possible job-analysis procedure. A multi-method approach was considered desirable despite its higher cost in terms of time and money. The two procedures were selected because they were felt to represent the best of all those available and because they complement each other. That is, each overcomes a potential weakness in the other.

The task inventory/CODAP method could be used alone. The function of the SME panels, i.e., identification of KSAs, could be accomplished by adding a KSA section to the task inventory booklet. Respondents could rate KSAs on individual 7-point scales relating to promotion and training. This procedure was not considered desirable for three reasons. First, the addition of a KSA section would have increased the inventory length to a point where the completion time for the inventory would be prohibitively long. Civilians cannot be required to complete the task inventory and every effort must be made to

secure their voluntary participation. As the inventory booklet was already quite long (31 pages), any additional length would have a negative effect on the return rate. Second, the identification of KSAs for promotion and training requires more information than can be obtained by a single rating scale on each topic. As is described later in this paper, three rating scales are used just to identify KSAs for promotion purposes. CODAP-type task inventories generally do not employ this number of rating scales. Third, a job analysis consisting solely of task-inventory data is not considered as defensible from a content validity standpoint as is one containing data from multiple methods. Items in a task inventory, like any other questionnaire, are subject to varying interpretations by respondents. No matter how carefully task statements are written, they will be perceived differently (i.e., erroneously) by some respondents. The problem of differential interpretation is especially acute in this project because one task inventory is being administered to employees in three different career fields. The potential for this problem may be greater for rating KSAs than for rating tasks. This is because employees are familiar with task statements in job descriptions and performance standards. They are not equally prepared to deal with KSAs. In general, the content validity of the job analysis will be strongest when the results of the task inventory are corroborated by another method.

The SME panel method also could be used alone. The function of the task inventory/CODAP analysis, i.e., identification and description of jobs, could be accomplished by having the panel members write task statements and rate them on the same scale(s) that would otherwise be used in the task inventory. In essence, this method would be a task inventory without a mailout. The same employees would provide task statements, rate them, and assign them to jobs. This procedure was not considered desirable for three reasons. First, it could have the appearance of a closed-door, "old boy" system to both employees and the courts. Even though it may provide identical results, the large-scale survey method appears more objective than the SME panel method. Second, it does not publicize the job analysis the way a large-scale survey does. The publicity function is important in this project because the promotion and training systems which depend on the job analysis results need to be accepted by employees in order to operate as desired. Employees may not pay much attention to newsletters, memorandums and other information-sharing attempts, but they must attend to a task inventory even if they choose not to complete one. A 100 percent sample is used in the task inventory mailout in part for its public relations value. Although published studies show that reliable results can be obtained with surprisingly small samples, a 100 percent mailout is seen as worth the relatively small extra cost because of the way it involves employees in the job analysis process. Also, the smaller the sample, the more the procedure resembles that of the SME panel methodology and the project desired two quite different procedures. Third and finally, an SME panel approach was not used by itself because SMEs were not comfortable with the task of defining all the jobs in the career programs. Management wanted the job analysis to determine "what is out there," or what is objectively found by the analyst to be present in the field. Management officials and good psychometric practice dictate that a few SMEs are poor candidates to attempt to specify the job tasks of several thousand employees throughout the world.

Qualitative-Quantitative Distinction

The task inventory and panel methods are considered by some to be at opposite ends of a quantitative-qualitative continuum. Our experience with these methods shows this not to be the case. We consider panels to be every bit as quantitative as the task inventory method (conversely, task inventories to be as qualitative as panels). A good deal of statistical analysis follows our panels. The main difference appears to be in terms of sample size. Perhaps the qualitative reputation of panels stems from the use of panels by some job analysts merely to provide narrative information on tasks or KSAs. In our method, panels are used to provide numerical ratings on a number of dimensions. To illustrate this, consider the steps the panel goes through in identifying KSAs for just one purpose, that of promotion:

1. Editing of important tasks. Panel members are provided with a listing of the tasks identified by the CODAP analysis as being most important for their job. ("Important" tasks are separated from all the tasks in the CODAP-produced job description by the use of some heuristic rule. One such rule typically used is that an important task is one that is performed by at least 50 percent of the respondents and is in the upper 50 percent in the cumulative group rating column of the job description.) The panel determines whether performance on those important tasks differentiates superior from average employees. The panel members do this by individually rating the extent to which superior and average employees differ on performance of each task. A three-point scale ranging from "not at all/slightly" to "significantly" is employed. Tasks on which there is little variation in job performance are dropped from further consideration. This step occurs because there is little predictive value in identifying KSAs for tasks that all employees perform equally well. Our ultimate goal is to identify KSAs that result in superior job performance.

2. Identification of KSAs and linkup with individual tasks. Some job analyses stop with the identification of job-related KSAs. Our procedure follows the identification of job-related KSAs with a rating of the KSAs differentiation ability. Panel members individually rate each KSA on the extent to which possession of it is important in distinguishing superior from average employees in performance of each differentiating (see step #1) task. A three-point scale ranging from "not at all/slightly important" to "highly important/critical" is employed. Each differentiating task is rated in turn. Identifying the KSAs that show the largest difference between superior and average employees is done because the most differentiating KSAs should be the most valid ones for selection purposes. For any job, there are many KSAs that are job-related. However, many job-related KSAs show only minor differences between superior and average employees and, therefore, are not very useful for selection purposes. Since there are practical limits on the number of KSAs upon which an applicant can be rated, the use of the most differentiating ones maximizes the validity of the rating process.

3. Linkup of KSAs with overall job. In this step, panel members identify the KSAs required for the job as a whole and determine how well those KSAs differentiate superior from average employees. This step is similar to the last one, the main difference being that the job-related and differentiating KSAs are identified for the job as a whole instead of for individual tasks. The overall-job linkup is used in addition to the task linkup because there is no accepted best way to combine KSAs ratings on individual tasks. Also, there may be KSAs required for superior overall job performance that are not related to individual tasks (e.g., the ability to handle numerous activities concurrently).

Perhaps the quantitative reputation of the task inventory method stems from the amount of computer time required and the quantity of printout produced by CODAP. There is no denying that the task inventory method requires a good deal of statistical analysis. However, there are many qualitative aspects of the task inventory method which are often overlooked or glossed over. Two general areas will be used to demonstrate the qualitative nature of the task inventory method. First, consider the process of determining the sampling plan, i.e., determining which employees to interview when writing task statements. This is the "Catch 22" of the task inventory method because the analyst has to have a pretty good idea of what the results will be before the job analysis is conducted. That is, one needs to have an idea of what is done at each location (major command, office, division, etc.) in order to interview the full range of employees and develop a comprehensive inventory. The interview plan is based primarily on qualitative information.

A second example of the qualitative nature of the task inventory method is the number of decisions necessary in the writing of task statements. The job analyst can alter the results of the CODAP analysis by altering the manner in which task statements are written. By writing more general statements, the analyst can "hide" differences between groups. Conversely, by writing very specific task statements, the analyst can make jobs appear more different. By judicious use of verbs that relate to level of responsibility (e.g., draft vs. write, propose vs. approve), the analyst can control whether grade and staff-operating differences appear in the CODAP analysis.

Thus, both job analysis methods have qualitative and quantitative aspects. The qualitative side of the task inventory method may be more apparent to those who have employed it repeatedly. The quantitative nature of the SME panel method may be hidden by the few analysts who continue to use it to produce only narrative information.

CONCLUSION

In order to meet current legal and federal guidelines in the implementation of Army programs, a rigorous job analysis is necessary. In the case of this logistics job analysis, we found that an integrated (multi-method, multi-purpose) approach was called for. Our experience is that using this multiple approach will provide the most useful product available.

Managerial Competencies Assessment for Army Civilians

Dr. Grenville C. King
U.S. Army Management Engineering
Training Activity

In the autumn of 1982, the U.S. Army Civilian Personnel Center tasked the Army Management Engineering Training Activity to develop a competency based system to support manager and executive development (Army Civilian Executive and Manager Development System, ACE&MDS).

Research on the project was initiated in January 1983 with the review and analysis of prior research studies. However, frequently during the course of the project, several major activities were initiated and researched concurrently. The multiple concurrent chronologies, involving the major design, development, and test activities of the ACE&MDS project are grouped as follows by functional category:

1. Isolating the specific managerial competencies which would be used in ACE&MDS.
2. Testing a competency assessment instrument for individual managers.
3. Identifying and evaluating training and development opportunities.
4. Developing an automated data base and processing system.
5. Designing and testing a complete managerial competency assessment program for Army management personnel at all levels and in all career fields.

The Managerial Competency Concept

The initial tasks that faced the ACE&MDS project team involved isolating the specific managerial competencies that would be used as the basic foundation for the system. An extensive literature search was conducted to review the development of the competency concept and to attempt to identify potentially useful assessment systems. Programs that had been developed within the Federal Government were given particularly careful review since their competencies were potentially more likely to be representative of the Army civilian population than studies of private-sector management. The most significant public-sector competency programs that were reviewed included: U.S. Army Training and Doctrine Command Soft Skills study, U.S. Air Force Training Command, U.S. Army Organizational Effectiveness Center and School, U.S. Office of Personnel Management(OPM) Management Training Needs Profile as modified for U.S. Army Tank Automotive Command (1976), and the OPM proto-type Management Excellence Inventory (1983).

An examination of all the foregoing studies, and numerous others of less significance indicated that there was continued interest in many circles in defining those competencies required to manage successfully. The various studies, however, approached the question of managerial competency

in different ways. Some were elaborations on the studies undertaken by Shartle at Ohio State University in the 1950's and 1960's; while others focused on the Roles of Management (Mintzberg, 1973) or investigated functional groupings of managerial work (Tornrow & Pinto, 1976). The definition of managerial competency was also quite diverse.

Definition of Competency

Early in the ACE&MDS project, the research team developed a standard definition of competency and adopted criteria that had to be met by a competency system and its assessment process for it to be of value supporting DA requirements. A Competency was defined for the purpose of ACE&MDS as, "Aggregates of behaviors and activities expressed in terms of knowledges, skills, and abilities, necessary to perform a task in an acceptable manner." A managerial competency was further defined as, "Aggregates of behaviors and activities, expressed as knowledges, skills, and abilities necessary to perform in a competent manner in the manager role."

A significant criterion of competencies for use in ACE&MDS was that the competencies must be based on observable, teachable, and trainable behaviors. Non-teachable aspects such as physical traits and characteristics were not used.

In addition to the teachable/trainable criterion, it was decided that a competency assessment system for the Army would have to have the following characteristics:

- o Accuracy; identifies true needs.
- o Objectivity; factual and unbiased.
- o "User-friendly"; easy to use; little time required to administer; timely; relevant feedback.
- o Relevancy; results would be job related and also allow information for career growth.
- o Hierarchical; data had to be in a form capable of easy consolidation for policy decision.
- o Cost effective; per capita processing costs had to be low enough to allow large-scale use.

Examination of Existing Managerial Competency Assessment System

When all the factors listed above were taken into consideration, three existing managerial competency assessment systems, two from the Office of Personnel Management, met enough of the criteria to be evaluated as possible systems for the Department of the Army.

After careful evaluation, it was determined that it would involve as much work to modify the existing inventory and test it as it would to build a new inventory. Further work with the OPM inventory was discontinued and the ACE&MDS project team started formulating a plan for a managerial competency assessment process tailored to Army needs.

The Army-specific competency problem which initially appeared as a major obstacle, was solved through the use of the competencies identified for the Automated Program for Executive Development(APED). This earlier AMETA project was performed for the SES Office, HQ, DA, in 1981.

In order to be beneficial for developmental purpose, the specific activities performed by Army executives had to be identified. Since the SES personnel came primarily from a broader management pool, the identification of managerial activities for APED was cascaded through all levels of management and expanded across the Department of Defense. Thus, the APED list not only applies for Army management, but also for Air Force and Navy management. The basis for APED's usefulness lay in its identification of 189 specific activities actually performed by Army (and DOD) managers as presented in the 1981 study, the APED activities were cast as teaching objectives grouped qualitively into 28 categories which were likewise grouped into nine major areas. After carefully examining the APED study and reviewing the study methodology with its authors, it was determined that with some refinement and testing, the study could be used as the initial basis for development of an Army-wide competency system.

Designing a Managerial Competency Assessment Inventory

Initial ideas on the ACE&MDS inventory format and content coalesced rather quickly. It was decided to base the inventory on the 28 competencies and nine Macro-Competency Areas instead of all 189 specific activities (which existed at that time) to avoid problems with length and time that had been observed with the OPM instrument. To offset potential problems with this approach, test subjects (i.e., individuals used to test the inventory instrument) were furnished with a list of all 189 activities to allow them to refer to specific activities when necessary.

Once the initial decisions were made on the number of competencies that would be used, the project team began development of the demographic portion of the inventory. An initial requirement established for ACE&MDS was that competency data would be able to be aggregated in various combinations for different groups of managers. The demographics would provide a mechanism that would allow the research team to not only aggregate data by various groups but also to directly address specfic characteristics of the population. As a result, data now exists to address questions of population mobility versus promotion, education, and experience factors in relation to grade, career field, etc. Since these questions can be addressed in multiples for different career fields, installations, level of management, and so forth, the demographics represent a valuable planning tool for personnel management.

While the demographic portion of the inventory was being designed, developmental effort was underway with the competency assessment portion of ACE&MDS. Since actual measurement of competency is so time consuming, it was determined the format of the ACE&MDS inventory would be a self-report based upon the concept that one person who best knows the strengths and weaknesses of that individual is that individual himself(herself). After using this self-assessment concept in tests with approximately 500 managers, it was found to be desirable to involve the immediate supervisor in the process.

To avoid similar problems observed with other inventories, the competency assessment part of the ACE&MDS inventory was segmented into three portions which were rated separately. In Part I, the respondent was asked

to rate the importance of each of the nine macro-competency areas to their job. This was a forced choice rank-ordering process where the person taking the inventory uses whole numbers, 1 to 9, without using the same number twice. In Part II, the respondent indicates for each of the 28 competencies on two factors, a. "Do you need training?" - YES/NO; and b. "How often do you use this competency?" - Never/Rarely/Occasionally/Frequently. Part III asks several questions that were generated from research with managers who were rated exceptional on their last performance appraisal. Each aspect of the competency assessment is independent thus avoiding problems of mixed quesion/answer results. The independence of these elements is maintained throughout the data summaries; however, they can also be considered together if the need arises. For example, the elements on an inventory can be combined to develop an individuals training needs priority in which case they are weighted and multiplied.

Tests with the Proto-Type Inventory

The ACE&MDS inventory underwent eight iterations, each of which were refinements of the previous version. All versions of the inventory shared the basic 3-part characteristic; therefore, the following discussion of tests with the inventory refers to the proto-type inventory as if it were one version.

The inventory tests were initiated in late March of 1983 with ten AMETA management students. Based on interviews with these students, the inventory was revised and tested with subsequent management classes. This process was repeated until, by late June 1983, 373 students had taken the inventory and been interviewed. The interviewing was done both individually and in groups and resulted in improved instructions and better formatting. Additional studies were also conducted, such as one performed with exceptional managers where inventory results were compared with the results of detailed critical incident interviews. These results were calculated manually. The reaction from the test participants was very positive, both to the inventory and to ACE&MDS as a system. A final test of the inventory involved test/retest reliability. This action was conducted to determine if the inventory gave stable results over time. The analysis of test/retest data indicated that the ACE&MDS inventory was highly reliable. With this information, it was decided to test the ACE&MDS inventory across a representative population of Army managers.

The Low-Scale Inventory Test Phase

The low-scale test of the inventory took place between July and September 1983. To insure a representative sample of participants, samples from installations in different geographic areas, from every Major Command(MACOM) and from all organizational levels within the commands, were selected. The result was a test population of 2,400 managers from 42 installations ranging in size from Tooele Army Depot, Tooele, Utah, to Headquarters Training and Doctrine Command(TRADOC), Fort Monroe, Virginia.

By this time, a rudimentary automated processing system was in place and participants were given the option to have training needs printouts returned to them; over 1,400 requested the training needs printouts. A

structured observation checklist was used by researchers to record the actions, questions, or comments of the participants as they participated in that test. As with the proto-type test, interviews were conducted of groups and individuals and information collected to improve the system. A long form inventory was also given at selected sites as a cross-check during this phase.

The outcome of the low-scale inventory test was that the ACE&MDS inventory worked. Modifications were made based on the information received from test participants but these concerned "user-friendliness" issues.

The Three-Step Inventory Process

As mentioned previously, the proto-type inventory relied totally on the self-report of the position incumbent. Analysis of the results of the proto-type test indicated that a small number of people were trying to manipulate results; for example, some reported training needs in everything, others reported no training needs at all. While the position incumbent is usually the best informed about the requirements of that position, a check was required to minimize this tendency of some managers to manipulate the inventory results. In addition to the personal manipulation problem, the study team was concerned about how the supervisor's perceptions could be considered in the process.

Even though the incumbent's perception of what his job demands are is likely to be very accurate, it is the supervisor who established the job requirements, evaluates performance, and initiates training requests to satisfy needs. The supervisor does this from his knowledge and perception of the incumbent's position. Hopefully, the two perceptions would converge; if not, a mechanism to reconcile the two views of job requirements and training need was needed. This mechanism evolved into the AMETA Three-Step Assessment Process.

In the Three-Step Assessment Process, Step 1 required that the subordinate fill out the ACE&MDS inventory based on how he views his job requirements and how he sees his personal need for training and development. In Step 2, the immediate supervisor fills out the inventory based on this assessment of the subordinate's job and on his assessment of the subordinate's need for training and development. Step 2 is completed without discussion between the two parties or to any reference as to what the subordinate wrote. In Step 3, the subordinate and supervisor meet and negotiate their perceptions of the job's requirements and the subordinate's need for training and development. The first two steps are important because they give both individuals the chance to record their views objectively without concern about the other person's perceptions.

Experiments with the Three-Step Process were conducted in August, 1983 at various installations. Matched pairs of actual subordinates and supervisors role-played the 3-step process. Observers watched the process and conducted post-test interviews to determine facility, accuracy, and reliability of the 3-step process. The superior-subordinate interaction that takes place during the Three-Steps also has important fringe benefits. It forces participants to compare perceptions using a standardized list of observable performance-based competencies; this puts that the interaction

on a more objective footing. Respondents in interviews reported that their interactions, while not always comfortable, proved enlightening since participants had previously relied heavily on assumptions regarding each others perceptions.

The Three-Step Process was incorporated into the ACE&MDS inventory. The inventory and 3-Step Process was now ready for a full system operation test.

To insure as accurate of a system test as possible, the system was tested throughout the management population of a continguous segment of a Major Command(MACOM). Everyone from team leader through executive; from camp/post/station through MACOM, HQ, would participate. Participants received individual training needs printouts. Summary data on organization populations were provided to local T&D officers after the test. Post test interviews were conducted with a sample of participants and will all T&D officers.

LOGAMP

The full ACE&MDS system was also used to support the Logistic and Acquisition Management Program(LOGAMP). The LOGAMP is a proto-type program to improve the managerial and technical competency development of high potential GS-13-15 personnel in a family of six Logistics-related career fields. While the ACE&MDS MACOM test was underway, the LOGAMP program was finalized. LOGAMP selectees were administered the ACE&MDS inventory as a means of developing IDP's. AMETA personnel provided follow-on assistance after the Three-Step Process by assisting the LOGAMP participants and advisors at a special IDP seminar.

The experience of the system operational test in the U.S. Army Corps of Engineers and the use of ACE&MDS to support LOGAMP clearly indicated all components of ACE&MDS were functional as designed. The total integrated system provides timely and accurate managerial competency needs identification for individuals and groups and is an effective mechanism to meet the developmental needs surfaced by the Three-Step Process.

TYPES AND QUALITY OF NATIONAL TRAINING CENTER DATA

Patrick J. Whitmarsh
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

The National Training Center (NTC) at Fort Irwin, CA provides U.S. Army heavy battalion task forces (TFs), controlling brigade headquarters and supporting units training with Multiple Integrated Laser Engagement System (MILES) and Live Fire under realistic combat conditions not found at home station. The NTC is in a position to provide Army decision makers information on battalion task force training, doctrine, and readiness in the context of the Army Training and Evaluation Program (ARTEP). One objective of the Army Research Institute Presidio of Monterey Field Unit (ARI-POM) NTC Training Research Program is to establish an operational, computer-supported NTC Data Management and Analysis System to support research on these complex issues.

The purpose of this paper is to present the types of NTC information currently available; research problems with respect to collection and merging of information; data quality issues; and a discussion of standardized procedures.

TYPES OF INFORMATION AVAILABLE

Operations Plan. The Operations Plan is a planning document provided by the NTC to the TF controlling brigade to be executed by the battalion TFs at the NTC. Information typical of the Operations Plan include the MISSIONS which will be executed, and the general Astronomical data across the time period included; a SCENARIO PLAN; a TIME/EVENT SCHEDULE for Mechanized and Armor TFs, and OPFOR; a GENERAL SITUATION narrative describing a hostile international condition with potential for war between the two superpowers: OPFOR and TF with GENERAL SITUATION UPDATE(S); an ANALYSIS OF AREA OPERATIONS narrative describing a General Description of The Area, Military Aspects of The Area, Effects of Characteristics of The Area; a DEFENSE INTELLIGENCE REPORT on the Handbook Krasnovian Army (OPFOR); and a series of OPERATION ORDERS (OPORD), WARNING ORDERS and accompanying overlay maps.

Take Home Package (THP). The THP is prepared for each TF Commander by the NTC Training Analysis and Feedback (TAF) Division. The document contains narrative and numerical descriptions of unit performance at the NTC and recommends additional training for home station training. THP information includes a General overview of the document in terms of Purpose, Scope and Organization; the Missions Conducted; a 14-Day AAR Briefing describing the Offense (missions) and Defense (missions) by Engagement Simulation in terms of Equipment Loss and Equipment Loss Ratio and by Live Fire in terms of Rounds Fired, Target Hits, Rounds Per Target Hits, Targets Hit, Rounds Per Targets Hit, Kills and Rounds Per Kill, and Trends describing TF Performance and Battalion Equivalent Loss Comparison; the Daily AAR Presentations describing in narrative each mission in the context of Operating System, Reported Event, Effect, Reason, Cost and Doctrine and describing numerically the Equipment Losses by Company/Team across Tanks, APC and TOW and the Radio Transmissions by Task Organization across Number of Transmissions, Average Length of Transmissions, Number of Transmissions Greater-Than-25-Seconds-Less-Than-50-

Seconds and Number of Transmissions Equal-To-and-Greater-Than-50-Seconds; and a List Of Audio/Visual Materials.

Company/Team Take Home Package. The Company/Team Take Home Package is prepared by the TAF Division for each TF company/team. The document describes in narrative form, the Plan, Prepare and Execute phases for each NTC mission for each Company/Team.

Video Tape After Action Review. The video tape After Action Review (AAR) is prepared by the TAF Division on each TF mission. It is intended to supplement the THP in establishing future home station training. The AAR, considered the key to training effectiveness, is a tactical discussion among all soldiers conducted by the senior trainer immediately following each mission training exercise.

Unit After Action Report. The Unit After Action Report is a detailed narrrative and numerical description provided by the brigade commander to the division commander on the NTC training period. The Unit After Action Report narrative includes a Narrative Overview, each TF Commander's Comments, Task Organization, Specific Comments and Recommdendations on Command and Control, Maneuver, Fire Support, Intelligence, Air Defense, Mobility and Counter-mobility, Combat Service Support, Survivability, and Recommendation for Improvement of the NTC Experience. The numerical contents include Maintenance Production and Major Assembly Usage, Vehicle Requirements, and Budget.

National Training Center Instrumentation System (NTC-IS) Data Tape. The NTC-IS is divided into the Range Data Measurement Subsystem (RDMS), Range Monitoring and Control Subsystem (RMCS), and Core Instrumentation Subsystem (CIS). The RDMS provides real-time position/location and engagement event data on all instrumented players during the TES. The RMCS monitors and controls all activities on the engagement simulation and live fire ranges. The CIS provides all real-time data processing and interactive play capabilities needed to monitor, command and control all engagement simulations. The NTC-IS Data Tape contains information replay of the battle for display on "NTC workstation" graphics terminals and also for input into the ARI-POM NTC Research and Training Data Base. The tape data are formatted for the INGRES relational data base management system (RDBMS) on the DEC-VAX 11/780 computer system. INGRES was selected based on the research data base (RDB) criteria: maximize data element relationality, user friendly, retrieval efficient, modularly expandable, allow researchers continuing study, reliably supported and meet computer system specifications. INGRES facilitates research by manipulating the existing 61 data tables. Continuing research anticipates restructuring, adding and deleting variables with INGRES.

RESEARCH PROBLEMS

Collection. Data collected from the battlefield must be interpreted from a given context. To get a complete picture for off-line analysis we need plans, scenarios, orders, communications, and accurate position location and firing data. In addition, we would like to have data on transient events such as smoke and on factors which are basically not instrumentable (e.g., what did the commander know?). This is an ideal not readily obtainable. Our goal is to work with NTC personnel to obtain as much as we can and understand what we are missing. This will permit us to better interpret performance.

Merging. Certainly one major issue is integration of information from multiple sources to permit more complete and effective analysis and interpretation. The ARI-POM NTC Research and Training Data Base is designed to accommodate text as well as numerical data thereby providing an efficient RDB.

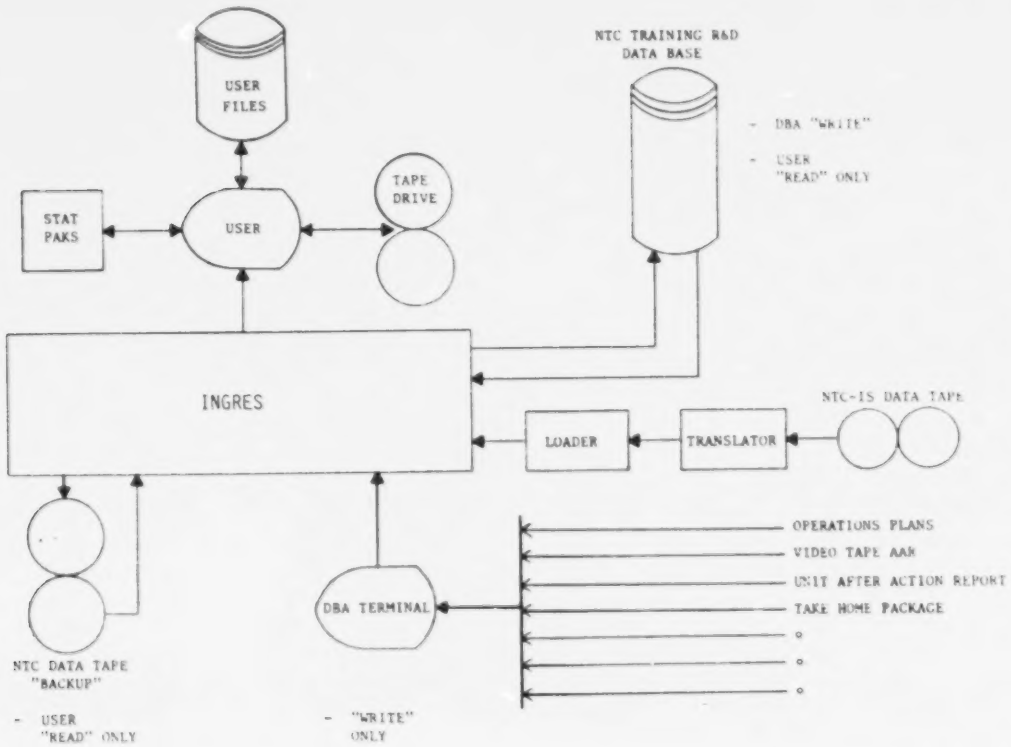


Figure 1
ARI-POM NTC Research and Training Data Base

Figure 1 indicates Operations Plan, THP, Company/Team Take Home Package, Video Tape After Action Review (convert audio to text) and Unit After Action Report are entered to the NTC Training and Research Data Base with INGRES by Data Base Administrator (DBA) from a terminal with "write only" privilege. The NTC-IS Data Tape is entered by tape drive thru NTC Translator and Loader software and INGRES by computer system operator. To simplify subsequent NTC-IS Data Tape operations, a NTC Data Tape "backup" is performed whereby a user can access the data tape thru INGRES with "read only" privilege. The user has access to the tape drive, statistical packages and user file system, and the RDB thru INGRES with "read only" priviledge.

DATA QUALITY ISSUES

Accuracy. Previous investigations have attempted to utilize the NTC data sources without knowledge of the conditions under which the data were collected, the result of which has been to limit the research to one data

source. With multiple sources and types of data and efficient merging, comparisons will be possible to identify data reliability. As a first step we investigated the accuracy of data entry on vehicle loss and radio communication variables across THP, INGRES RDB, Graphics Displays and Summary Statistics derived from the NTC-IS Data Tape for seven selected TF missions. The results indicated perfect agreement, in all cases, between the Summary Statistics and the INGRES RDB only. Some of the problems are attributed to missing data entries by operator personnel and possible software errors.

STANDARDIZED PROCEDURES

Operational Control Units. Data quality can be improved by use of standardized procedures, at the same time recognizing data quality can never be perfect because some data are collected during real time. However, procedures can be designed to support both training at NTC and improve data utility collected for off-line analysis. We, at this time, are preparing a procedural document to assist the NTC CIS Operations.

CONCLUSION

The NTC represents a complex environment providing the Army, for the first time, an opportunity to evaluate training readiness of TF units under simulated combat conditions. This paper examined types of information currently available, identified some research problems, discussed data quality issues and standardized procedures for the purpose of supporting the NTC RDB. A fully supported NTC RDB is the key toward effectively analyzing unit performance, thereby providing for an efficient and effective fighting force.

REFERENCES

Banks, J. H. (1985, May). Some issues and concepts for unit training and evaluation at the National Training Center and at home station. Paper on evaluation of collective training for technical panel UTP-2 of The Technical Cooperation Program (TTCP). Alexandria, VA: U. S. Army Research Institute.

Kroger, A. (1983, April). Considerations in the establishment and use of a research data base for analysis at Army training (Jet Propulsion Laboratory Working Paper). Pasadena, CA: Jet Propulsion Laboratory.

Nichols, J. J. (in preparation). The DeAnza Primer (BDM Research Product). Monterey, CA: The BDM Corporation.

Science Applications Inc. (1983, June). EMC/TAF operating manual for the NTC Core Instrumentation Subsystem (CIS) (500 player system) (Subcontract No. SD-6200). La Jolla, CA: Science Applications Inc.

Science Applications Inc. (1984, June). INGRES table formats and descriptions for a prototype National Training Center (NTC) research data base system. La Jolla, CA: Science Applications Inc.

Whitmarsh, P. J., & hamza, A. N. (in preparation). National Training Center data reliability: A comparison of multiple data sources (ARI Technical Report). Alexandria, VA: U. S. Army Research Institute.

An Overview of ARI's Research Program on the National Training Center

James H. Banks
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

The National Training Center (NTC) was established at Fort Irwin, CA to train battalion task forces under highly realistic and intense conditions which are not obtainable at home station. A second purpose was to provide information to help the Army evaluate its training, doctrine, organization, equipment, and readiness. Thus, the NTC provides capstone training and potentially permits measurement of the output of the Army unit training system. However, combined arms training exercises of the scale, complexity, and realism of those at the NTC have never before been conducted, nor has information of this richness been available for analysis and interpretation. ARI's NTC-based research program was established to help get the maximum benefits from the NTC and to increase the effectiveness of ARI's overall R&D program. I am going to provide an overview of the program, outline a few of the problems in measurement and interpretation of unit performance, and in diagnosing and correcting problems when they are detected, and generally describe what we hope to do. Other speakers will describe the NTC, the types of data available and potentially available, and some early analyses.

Measurement and Interpretation of Task Force Performance

The NTC provides an unparalleled opportunity to objectively measure and analyze unit performance to detect strengths and weaknesses, typical performance, performance ceilings, and trends. Measurement and interpretation must take into account both the nature of the Task Force and the combat environment in which it must function.

The Battalion Task Force is a complex system containing maneuver, intelligence, fire support, air defense, mobility/countermobility, and combat service support elements or subsystems, all bound together by a command and control subsystem, as shown in Figure 1.

916 1009

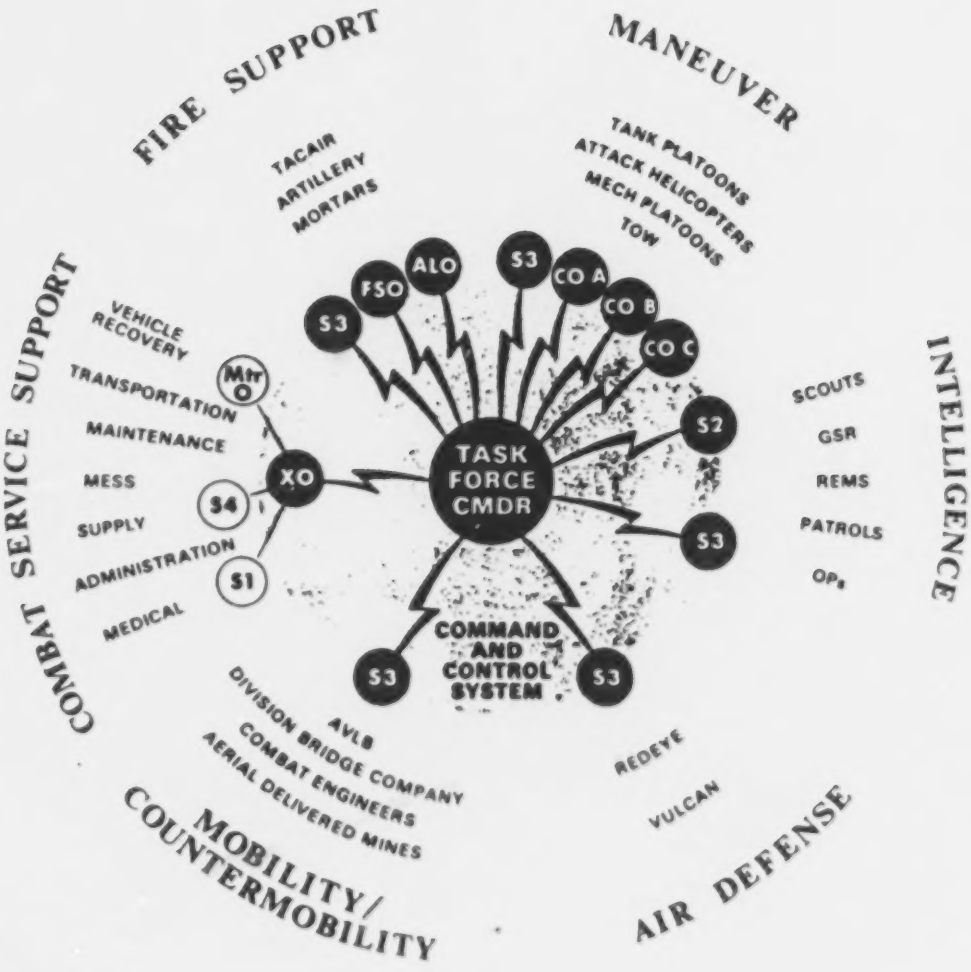


Figure 1
One Model of the Battalion Task Force as a System

The combat environment for the Task Force--or the NTC as a high-quality simulation--is complex, rapidly changing, and uncertain. "Correct performance" always involves trade-offs of capabilities, opportunities, and risks, and always in the context of the commander's orders and intent for a mission. Moreover, friendly actions are opposed and countered by an intelligent adversary. Thus the problems faced in combat, by the leader and by the soldier, are typically "fuzzy"--not clearly stated, where the needed information is not all available, where no set procedure can be used to reliably produce an answer, and where there may be no single answer that is demonstrably correct.

Given the complex nature of the Task Force and the combat environment and requirements, the problems in measuring and interpreting unit performance at the NTC are twofold. First, facts do not "speak for themselves": rather, they are meaningful only in relation to other facts and expectations. These relationships can be informal (e.g., "common sense") or quite elaborate, formal, and explicit, as in scientific theories. For the major relevant problem areas--how to fight, how to evaluate collective training and performance, how humans learn, how to instruct teams--the theories are not well developed or integrated. We,therefore, need descriptive models of the Task Force system and subsystems to guide observation and interpretation. Initially these can be derived from doctrine but they need expansion, synthesis, and empirical validation. Second, enormous practical problems exist in operationally defining concepts in terms of procedures and data elements, from the NTC or elsewhere, and actually conducting the desired analyses. While "quality control" of NTC instrumented data can certainly be improved, it will never be perfect because it is collected during real training exercises. In addition, many aspects of performance which might be desirable to observe are not readily collectable by automated methods, and capabilities for non-automated collection are limited and hard to reliably and validly implement. Under these conditions, analyses must be highly robust i.e., permit interpretation despite data gaps and complex interactions. Traditional data analysis methods are not well suited for such applications. However, human beings do, routinely, process and use information of this type. Therefore, we will be interested in developing "expert" approaches and models to supplement conventional methods.

The NTC as Part of the Unit Performance System

The NTC is not a stand-alone activity but, rather, is part of the Army training system. Interpretation and correction of performance observed at the NTC requires not only measurement of performance but also understanding of the effects and interactions of inputs to unit performance.

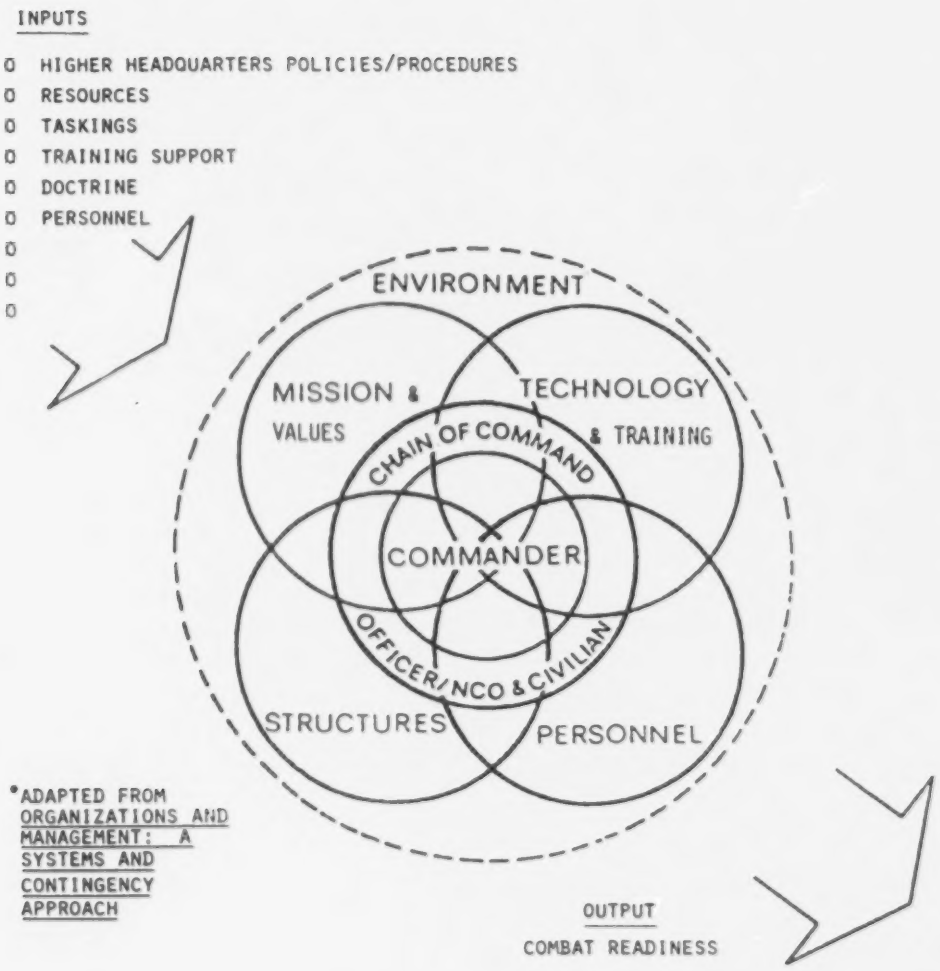


Figure 2
A Model of Unit Performance

This model views the unit as an open system composed of subsystems--missions and values, technology and training, personnel, structure, leadership. Thus, factors internal to the unit--training validity, timing, completeness, and mastery level; leadership priorities; individual and organizational values; morale; cohesiveness; etc.--all are in active and continual interaction with each other and with the environment in which the unit operates. Inputs include those found in Table 1.

Table 1
Inputs to Units

Doctrine	Threat analysis, strategy, missions, tactics, TO&E.
Institutional Training	Validity, timing, completeness, mastery level, training evaluation, performance accountability, values, quality of exported training support products.
Unit Training	Training resources, training evaluation, performance accountability, quality of aids and literature, fill level/job assignments/MOS mismatch, force modernization/transition, housing and post services, management skill, values, leadership.
Personnel	Recruiting, selection, classification, job design, aptitude measurement, school assignment, unit assignment, promotion, retention, elimination, retirement, pay and benefits.
Equipment	Operability, maintainability, durability/reliability, trainability, performance capability, cost.
Logistics	Speed and range, quantity/lift, survivability, economy/consumption, cost.
Societal Values	Attitudes towards the military, national goals and pride, etc.

To obtain maximum benefits from the NTC, it must be treated as a part of the unit performance system. Therefore, interpretation and improvement of performance must consider both home station factors and outside inputs to the unit. ARI's R&D concept takes this into account.

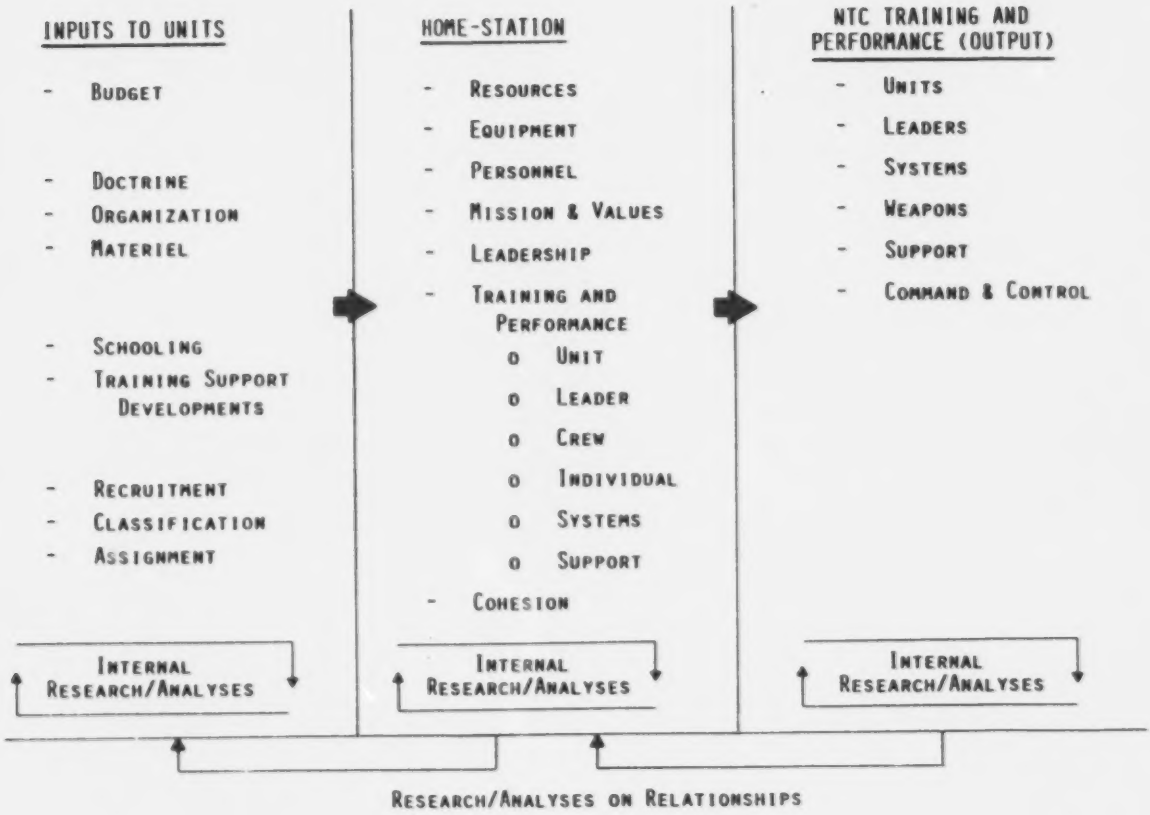


Figure 3
Research and Development Concept for Unit Performance System

In this concept, research on NTC training and performance, on home station factors, and on the various inputs to units will be conducted because of their importance in their own right to Army sponsors and clients of the ARI program. In addition, ARI has established the Unit Performance R&D Center at its Presidio of Monterey Field Unit, with scientific and Army training expertise and computer analytic and data base capabilities for handling data from NTC and other sources. As the data base is established over time, it will permit development of a scientific and practical approach to the measurement and interpretation of unit performance and, for the first time, effective research on the relationships and interactions of factors in the unit performance system. The primary sponsor for the NTC-based program is the Combined Arms Training Activity which is the Army executive agent for use of NTC information for assessment and improvement of Army training, training support, doctrine, organization, and materiel. It is expected that research at the Center will also potentiate the effectiveness and value of other ARI research and development.

LEADER PERFORMANCE CRITERIA AT THE NATIONAL TRAINING CENTER (NTC)

Earl C. Pence
U.S. Army Research Institute

Introduction

The presentations you have heard thus far have described the operation of the NTC and the types of data which are currently generated during the training exercises. For the most part, these data describe WHAT happened during the preparation and execution of a mission and are performance data at the unit level (battalion, company or platoon). The focus of the current presentation is on the development of a new data collection system which will provide information to aid in understanding WHY certain events or outcomes occur at the NTC. The new data collection system is being developed to aid the NTC Observer/Controllers (OCs) in observing and recording performance indicators of leaders at the battalion, company, and platoon levels.

The Current System

The OCs at NTC have, of course, been observing the actions of leaders at NTC since the time the training center first opened. In fact, the After Action Review (AAR) process used to provide feedback after the end of missions at NTC is a leader-oriented process in that the information at the battalion and company levels is provided to the key leaders in the battalion. The information itself, however, is focused on unit performance with little emphasis on leader performance except for probing questions as to why leaders made certain decisions or took particular actions which the OCs judged as critical to the outcome of the mission.

In August of 1984, the Leadership and Management Technical Area of ARI, in conjunction with the Center for Army Leadership (CAL) at Ft. Leavenworth, began work on a leadership research effort at the NTC. The initial task was to assess existing leadership training and data collection processes at NTC and to determine the potential for enhancing the quantity and quality of leadership data collected at NTC as well as the means for improving the development of leaders participating in exercises at NTC. During the fall of 1984 a research team composed of a researcher from ARI and a Major in the research branch of CAL spent a two week rotation in the field at NTC. The team traveled with the OCs and focused their observations on the manner in which the OCs gathered information, recorded observations, and provided feedback to leaders training at NTC. The research team also recorded the conditions under which the OCs performed their work.

While I will describe the conclusions drawn from these observations verbally, I think a few slides illustrating the conditions under which the OCs perform their job will provide a better understanding than any verbal description I could give.

The first three slides illustrate the NTC OC's primary work site--the 1/4 ton jeep. The jeeps are loaded with 7 to 9 days' supplies and all the OC's tools of the trade, including a lot of pyrotechnics and a map strapped to the hood. The fourth slide illustrates an example of an OC observing the deliv-

---

The views expressed in this paper are those of the author and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

ery of a company operations order. The fifth through seventh slides give you an idea of what the OCs observe while chasing tanks and APCs (Armored Personnel Carriers) during a training mission. The eighth slide illustrates another part of the OC's job, crawling up on tanks to verify MILES kill codes. The ninth and tenth slides illustrate the platoon OCs meetings with the company OC at end-of-mission to compare notes and prepare the company level AAR and the delivery of the company AAR. The eleventh and twelfth slides illustrate the preparation process for the battalion AAR and the cite for the battalion AAR. The final two slides illustrate the typical sleeping conditions for the OCs when they have the opportunity to obtain more than an hour or two of sleep. If they have only a short time (which is typical) they will simply sleep sitting in the jeep.

Initial Research Findings

The key findings from the research team's initial visit to the NTC are listed on the first overhead slide.

SLIDE 1

INITIAL RESEARCH FINDINGS

The observer/controllers receive little or no training.

Considerable variation between OCs on decision rules guiding observations and feedback.

OCs made notes of their observations in a variety of means:

* 3 x 5 note cards
* Small memo pads
* In grease pencil on map cover on hood of jeep
* In grease pencil on top of ammo can

Notes used primarily for AARs and relatively little of the information was captured permanently.

A second round of observations in the field at NTC by the same research team in June of 1985, followed by interviews with approximately 30 past and current NTC OCs, confirmed all of the findings noted above. The only major new finding from the second data collection effort was that the company OCs are now required to produce a company level Take-Home Package which makes it even more critical for them to record their observations during or immediately after each mission.

The observations in the field and the interviews with the OCs indicated the need and opportunity for enhancing both leader performance data collection and leader development at NTC. The primary means for accomplishing both goals will be to develop a system which encourages more systematic, consistent, and accurate observation and recording of leader performance followed by feedback in the AAR on key leader events impacting on unit performance. The system must meet the criteria contained on the next overhead slide to be acceptable to the OCs who would use it.

SLIDE 2

CRITERIA FOR SYSTEM ACCEPTABILITY

1. The system must not increase OC work load.

2. The system must be convenient to use under extremely harsh environmental conditions.

3. The system must provide information which can be used almost immediately for preparing and delivering AARs.

The Design of the System

There are actually two distinct challenges to be met in designing the data collection system for leader performance criteria at NTC. The first is to develop the appropriate content for the data collection system (i.e., identify performance dimensions and operational measures of the performance dimensions). The second challenge will be to design the format or technology which provides a data collection process acceptable and useful to the NTC OCs.

Content

Development of the actual performance dimensions will be accomplished using a modified BARS (Behaviorally Anchored Rating Scale) development procedure. The key steps in this process are listed on the next slide.

SLIDE 3

STEPS IN DEVELOPMENT OF CONTENT FOR LEADER PERFORMANCE MEASURES

1. OC interviews to identify potential performance dimensions and examples of effective and ineffective leader performance.

2. Development of performance dimensions with definitions.

3. Work with OCs to develop behavioral anchors, performance indicators and decision rules.

4. Formative evaluation of anchors and performance indicators.

5. Development of OC training program.

6. Validation of the leader performance measures using unit performance measures as criteria.

The performance indicators and decision rules referred to in steps 3 and 4 will be guidelines to be used by the OCs in determining the appropriate rating to be given on any particular performance dimensions. For example, on

a performance dimension related to communication by the leader, the performance indicators and decision rules would probably include a series of questions which the OC could ask subordinate leaders and soldiers to determine the extent to which mission-related information has been communicated. The leader's rating on the communication dimensions would depend, in part, on the answers the OCs received when they asked the questions. The use of such performance indicators and decision rules combined with guidelines on recording specific leader behaviors would be part of the OC training program.

Format or Technology

The most difficult challenge for the data collection system is to package the system using a format or technology which meets the criteria regarding no increase in OC work load, convenience of use, and immediate information availability. The most promising concept for meeting these needs involves using an electronic clipboard under development by ARI's Training Research Laboratory. The clipboard is a small, hand-carried device which makes use of a touch sensitive screen for recording data. The format of the screen is programmable and seems ideal for a menu-driven data recording system. The menus operating on the screen could provide branching which would enable fairly detailed recording of observations or ratings with only two to four touches to the screen. The system would require no key punching and would be considerably more efficient and convenient to use than traditional check lists. The figure on the screen now illustrates an example of how such a menu-driven data collection system might work.

SLIDE 4

MENU-DRIVEN DATA RECORDING SYSTEM

The output of data or information from the clipboard could take any of a number of formats. The data could be presented on the clipboard screen, it could be printed, or it could be transferred directly to a computer. This variety of options would meet the need for immediate feedback of information in the field as well as eliminating the need for handling paper documents in the transfer of data for research purposes. Specific observations recorded on pre-printed, color-coded 3x5 note cards could be used to supplement the data collected on the automated system for illustrative purposes to explain ratings and provide feedback during the AAR.

Progress on System Development

Content

Interviews with the NTC OCs were conducted in June of 1985 and the initial content analysis of the interview transcripts has been completed. Examples of the types of leader performance information obtained in the interviews are illustrated on the next slide.

SLIDE 5

CHARACTERISTICS OF EFFECTIVE LEADERS AT NTC

PLANNING AND UTILIZATION OF TIME

Prepares detailed plans and gives specific guidance on what he wants to happen. (Communicates plans well, gives complete and cleare operations orders.)

Conducts planning suich that his subordiantes have time to prepare for the next mission (manages time well).

Uses time effectively and sets work priorities.

COMMUNICATION OF INTENT

Communicates his intent as a commander clearly.

Disseminates information to the lowest level possible and keeps people informed.

Understands his commander's intent.

SUPERVISION AND STANDARDS

Knows what he wants to see on the ground (knows the standard).

Communicates what he wants with authority (communicates standards).

Enforces standards and hold individuals accountable for doing their jobs (makes rapid corrections).

Supervises tasks after giving order (i.e., checks progress, walks the line, inspects positions or ensures this is done).

This information will be combined with previous research findings to develop an initial set of approximately 12 leader performance dimensions which will be taken to NTC for modification and operationalization by the OCs.

Format

The initial development of the electronic clipboard is nearing completion and a number of the devices will be available for pilot-testing by ARI during the current fiscal year. The initial pilot testing of the clipboards will occur at a site other than NTC. Preliminary work on a format that may serve as the basis for designing the menu-driven system to be programmed into the clipboard is currently underway.

TITLE: Battalion Performance on the Live Fire Range at the National Training Center (NTC)

AUTHORS: Thomas K. Forsythe and William J. Doherty

CORPORATE AFFILIATION: The BDM Corporation

Introduction

The training benefits derived by units training at the National Training Center (NTC) have been of considerable interest to the U.S. Army. The Army Research Institute (ARI) has developed and initiated a programmatic research effort to assess these benefits. As a preliminary step in this effort, an exploratory data analysis of the performances of battalion task forces on the live-fire range at the NTC was conducted. As stated by Tukey (1977) the purpose of exploratory data analysis is to be detective in nature not confirmatory. Thus, the investigation reflects incursions into the data designed to explicate the structure of the data rather than to confirm a particular model of the data. Using the data provided in 54 Take Home Packages for the period of early 1982 to late 1984, it was possible to examine battalion performance for three live-fire scenarios: Defend from a Battle Position (Day), Defend from a Battle Position (Night), and Movement to Contact (Day). The presentation of these data was organized around three primary issues:

(1) Has battalion performance changed over time at the NTC?
(2) How do the performances of the Armor and Mechanized Infantry Task Forces differ?
(3) What factors seem to be related to performance at the NTC?

Data Source

The live-fire data were extracted from fifty-four (54) Take Home Packages (THP). These data represented the live-fire performances of 96% of the battalion task forces that trained at the NTC from early 1982 to late 1984. Data from the remaining 4% (2 battalions) of the task forces were either not available, or erroneous due to target equipment malfunctions on the live-fire range. It was felt that exclusion of this small amount of data from the investigation would not adversely affect its generalizability or statistical power.

The live-fire results reported in the THP generally included the following data.

- Number of Targets
- Percent of Targets Killed
- Tank Rounds Fired
- Tank Round Hits and Kills
- Tow/Dragon/Viper Laser Firings
- Tow/Dragon/Viper Laser Hits and Kills
- Number of Each Friendly Weapon System

These data presented ample opportunities for conducting preliminary research on battalion task force performances at the NTC. The performance data was first extracted from the Take Home Packages and a computer data base was established to assist in sorting and analyzing the data. In all cases, unit designations were omitted to preserve unit anonymity.

Data Analysis

As indicated earlier, an exploratory data analysis (Tukey, 1977) approach was employed in this study. This resulted in a series of analyses being conducted and all aimed at understanding the structure of the data from the live-fire range at the NTC. Specifically, analyses were aimed at satisfying the first two objectives of the exploratory data analysis approach (Hartwig and Dearing, 1979):

(1) Understand each variable as a separate entity.
(2) Understand pairs of variables as relationships.

A number of data analytic techniques were applied to the live-fire data base. The initial efforts used univariate descriptive statistical techniques. The results of these were used to generate a picture of performance at the NTC. Generally, the results were transferred into graphic display. Somewhat more sophisticated techniques including T-tests and regression were used to examine bivariate relationships between different factors and performance at the NTC.

An important decision influencing the results contained in this report was the selection of a primary variable as measure of task force performance. Of the available variables, the one which seemed to most directly reflect overall unit performance was the PERCENTAGE OF ENEMY TARGETS KILLED.

It was also decided that because this report was designed to provide some early insights into unit performance changes at the NTC, only individual battalion performances would be considered and analyzed. That is, their performances would not be aggregated and analyzed at the brigade and division levels. However, an analysis of that order would be a logical follow-on to this analysis once a better understanding of the performance data at the battalion level is developed. For the same reason, only the effects of the tank ballistics are considered in this study although the performance of the Tows/Dragon/Viper laser firings on the live-fire range might also be investigated for a subsequent study. (It must be noted, however, that the TOW/DRAGON/VIPER performance analysis would not be as meaningful as the tank performance analysis due to the narrow range in values for the laser firings as compared to the tank ballistics data.)

Results

Using the "percentage of target kills" data as an indicator of meaningful unit performance on the live-fire range at the NTC and comparing the battalion performances in the first 18 months with those of the next 12 months of a 2-1/2 year period (early 1982 to late 1984), this study found that:

(1) The percent of targets killed by the tanks of the Armor and Mechanized Infantry Task Forces increased between the first and second periods of this study.

(2) The increase in percent of targets killed between the first and second periods was attributable to both one-time visitors to the NTC and to repeat visitors to the NTC. Therefore, the differences in performance was a function of some phenomena associated with time and not just some advantage factor acquired through repeat visits to the NTC as might be expected.

(3) The change in performances on the live-fire ranges was probably not a function of the activities involved in the operation of the live-fire range by NTC cadre; i.e., the live-fire exercises were conducted in a uniformly consistent manner throughout the 2-1/2 year period of this study.

(4) The change in percent of target-kills over time was not related to gunnery accuracy as this did not change over time. However, a significant increase in the volume of tank rounds fired by both the Armor and Mechanized Infantry Task Forces in the second period was likely related to the observed increase in target kills.

(5) The increase in rounds fired was related to an increase in number of tanks assigned to the task forces (particularly to the mechanized units) in the second period and to an increase in the number of rounds fired per tank.

(6) A positive and statistically significant relationship was found between the number of rounds fired per tank and the percent of enemy killed. This relationship was strongest for the day attack and day defend missions.

References

Hartwig, F. with Dearing, B.E. Exploratory Data Analysis. Beverly Hills, CA: Sage, Inc. 1979.

Tukey, J. Exploratory Data Analysis. Reading, MA: Addison-Wesley, Inc. 1977.

TITLE: Analysis of NTC Force-on-Force Performance
AUTHOR: Judith J. Nichols and William J. Doherty
AFFILIATION: The BDM Corporation

Introduction

U.S. Army battalions have been training at the National Training Center (NTC) at Ft. Irwin, California since 1981. The objective of NTC training is to provide a facility where units can undergo essential combined arms training that cannot be accomplished at home stations due to physical limitations and the prohibitive cost of providing realistic training environment. A secondary objective of NTC training is to gather information which can be used to contribute to the improvement of doctrine, tactics, training systems, equipment and procedures in the U.S. Army.

In support of the NTC information-gathering objective, the Army Research Institute (ARI) has developed a research program which includes as a technical objective the collection, organization, and analysis of NTC data in order to assess those training benefits which may be accrued by the units training at the NTC. This report is an investigation of the force-on-force data that are contained in the Take Home Packages (THPs) which are compiled and issued to the training units at the end of each NTC rotation. The focus of this study was the relationship between task force performance on the NTC battlefield and the force modernization (i.e., the changeover from H- to J-MTOE Series organization) which took place during the period under study.

Background

The NTC concept is geared to the training of the battalion task force (TF). Each battalion task force participates in approximately six force-on-force exercises during a two-week rotation period at the NTC. These exercises usually are more or less evenly divided between offensive and defensive operations using laser-based engagement simulation instrumentation to provide real-time casualty assessment. The simulator, the Multiple Integrated Laser Engagement System (MILES), is used on all principal weapons and casualties are assessed when a weapon fires and the MILES laser hits a target. In addition to force-on-force training, units also perform three missions on the live fire range during their rotation. The results of the investigation of live fire performance are presented in a separate paper for this symposium.

The scenario dictates the force ratios of the combatants. While terrain and scenario options are limited, no two scenarios are exactly the same. When the TF conducts defensive missions

they are always attacked by an OPFOR that replicates a Motorized Rifle Regiment. When task forces conduct offensive operations, they originally encountered a defending Motorized Rifle Company. However, in the summer 1984, the force ratio was changed to deploy a defending Motorized Rifle Battalion (-). Analysis of the effect of this change resulted in no significant differences, so the data was pooled across time periods.

Scope

Sample

The sixty-four (64) battalions which underwent training at the NTC during the period February 1982 through January 1985 are represented in this analysis. The 64 battalions represent the rotations from six divisons and two separate brigades located in the Continental United States (CONUS). The 64 battalions included 32 armor and 32 mechanized infantry units which were cross reinforced to form 64 combined arms task forces (i.e., 32 armor heavy and 32 mech heavy TFs). Two rotations by the opposing force (OPFOR) -- i.e., two MTOE battalions permanently stationed at Ft. Irwin -- were also included in the sample. The OPFOR battalions did not undergo the standard rotation series of exercises but rather performed a mini-ARTEP. However, the data extracted for the two missions performed by the OPFOR battalions (i.e., movement to contact and defend in sector) were not substantially different from the data of the rest of the sample and so were included in the analysis.

Data Sources

As previously mentioned, Take Home Packages (THPs) are prepared and issued to the training units at the termination of each NTC rotation. Separate packages are prepared for the respective armor and mechanized infantry task forces. The THP is an overall description of unit performance during the rotation and includes statements of performance trends during the 14-day rotation period. The THP is the final compilation of all after action review scripts and encompasses an assessment of all seven operating systems, live fire gunnery data, and TF and opposing force (OPFOR) aggregate losses.

The force-on-force results are reported in the THPs in several formats and include the following data:

- TF and OPFOR vehicle loss summaries for Offensive Engagement Simulation (OES) operations

- TF and OPFOR vehicle loss summaries for Defensive Engagement Simulation (DES) operations

- Numbers of TF vehicles started and killed for each mission

- Numbers of OPFOR vehicles killed for each mission

Offensive and defensive operations summary data were extracted from the 64 THPs. Mission-specific data for the six most commonly performed missions (i.e., movement to contact/meeting engagement, deliberate day attack, deliberate night attack, defend in sector, delay in sector, and defend from a battle position) were also extracted from the THPs and combined with the summary data to construct the data base for this analysis. Each individual task force was coded to preserve anonymity and the data were then subjected to a number of statistical operations.

Data Analysis

The task force was designated as the unit of measurement and the data were divided into two groups in order to identify differences in task force performance at the NTC during the period that force modernization was taking place. The first group included all H-MTOE Series rotations occurring between February 1982 and January 1985 and the second group included all J-MTOE Series rotations occurring during the same time period. The two groups were further refined to identify differences between mechanized infantry and armor TFs wthin the H- and J-Series categories.

The limited scope of force-on-force data available in the THPs severely constrained the analysis in several areas. For example, information on the numbers of OPFOR tanks and APCs starting each mission was not available in the THP thus prohibiting the derivation of percent of enemy vehicles killed during an exercise. Although the force strengths for Motorized Rifle Regiment and Motorized Rifle Company are available in FM 72-1, The Tank and Mechanized Infantry Battalion Task Force, Appendix H, this does not allow for instances of vehicle breakdown in the field, numbers of vehicles actually available to the OPFOR during specific individual missions, and so forth. For the purposes of this report, it was decided that the assumption of consistent standards for OPFOR strengths would not be made.

Two measures of battalion task force performance were selected for study. The first, "%Lost", represents the percentage of TF vehicles lost on the NTC battlefield -- i.e., numbers of TF vehicles killed at each mission / numbers of TF vehicles started

at each mission. This measure was available for each TF vehicle type killed (i.e., APCs to include TOWs, and tanks) as well as an aggregate value ("Total Vehicles Lost") that combined all vehicle types. Results of preliminary analyses revealed that TOW losses were not obviously different from APC losses and so TOW data were incorporated into the APC data to produce an APC/TOW category.

The second dependent measure, "Casualty Exchange Ratio", consists of the ratio of vehicle casualties for the two forces -- i.e., numbers of TF vehicles killed / number of OPFOR vehicles killed -- and was calculated for both specific vehicle types (i.e., APC/TOW and tanks) and across vehicle types.

Results

Analysis of the percentage of task force lost indicated little difference between H-series and J-series. The only statistically significant difference occured for the percentage of APC/TOWs lost on the Offensive missions. The direction of this difference indicated that J-series task forces lost a lower percentage of APC/TOWs on the average than did H-series task forces (13% fewer APC/TOWs were lost by J-Series task forces -- significant at the $p<.01$ level). The difference was found for both Armor and Mechanized Infantry task forces, though the size of the difference was greater for the Mechanized Infantry (18% fewer APC/TOWs lost by J-Series Mechanized Infantry task forces compared to 14% fewer APC/TOWs lost by J-Series Armor task forces significant at the $p<.01$ level). No statistically significant differences occured in the contrasts between H-series and J-series on the Defensive mission summaries.

The lack of a clear trend in performance differences between H and J-series units precludes any inferences being drawn concerning the force-multiplier effects of reorganization under the J-series MTOE.

Analysis of the casualty exchange ratios again produced significant differences only for the Offensive missions. The contrast between H-series and J-series showed no statistically significant differences for either Armor or Mechanized Infantry task forces. This was true regardless of mission type. However, there was a statistically significant difference between the two types of task forces in the exchange ratio for APC/TOWs on Offensive missions. The mechanized infantry task forces under H-series had consistantly greater casualty exchange ratios than the Armor task forces (H-Series mechanized infantry task forces averaged two (2) more combat vehicle losses per kill in offensive operations than did H-Series armor units -- significant at $p<.05$). Analysis of the casualty exchange ratios also showed considerable

variability in performance across the six mission types examined in the study, with the greatest fluctuations found in the Movement to Contact mission.

Based on the casualty exchange ratios, the OPFOR out performs the battalion task forces by achieving a much higher kill ratio of tanks and APC/TOWs when on both the offense and defense as compared to the TFs when they are on the offense or defense.

Discussion

The results of the analyses indicate that the determinants of force-on-force performance are more complex than the current information from Take Home Packages allow. Considerable variation in battalion task force performance was found in the results reported above. Yet, only a small portion of that variation seemed to be related to diffrences in the organization of task forces (H-Series vs. J-Series) or to the type of task force (armor vs. mechanized infantry). Further, the differences in performance along these dimensions were not consistent either in terms of type of mission (offensive or defensive) or type of performance measure (percentage lost or casualty exchange ratio). This lack of systematic results and the relatively small size of the relationships provide considerable evidence that force-on-force performance can only be understood in the full context in which that performance took place. To accomplish that end, it will be necessary to employ the full range of NTC digital data and possily supporting data such as the commo tapes.

Despite the limitations of the results, several findings are of note. There was considerable variation in performance of the task forces measured either in terms of percentage of vehicles lost or by the casualty exchange ratio. Variation in performance was greater for offensive missions than for defensive missions. This suggests that either the causal mechanisms for offensive performance are more complex than for their defensive counterparts or that training in the elements of defense are more universally applied and incorporated.

Another finding of note is that performance variation was least visible in the loss of tanks. Differences between type of task force and task force organization, when tested on the percentage of tanks lost or the causlaty exchange ratio for tanks, failed to produce any significant differences. As tanks constitute the primary weapon system on the NTC battlefield, performance characterisitics in this area are particularly important. The high level of loss in this area and the lack of

relationship to the analysis variables studied here indicate a more generic cause for the troubled performance. The nature of this cause needs to be studied.

The last significant finding from these results concerns the performance of the reorganized task forces. (J-Series). Since performance for these task forces was not found to be consistently better than their earlier counterpart (H-Series), it is important to determine why this should be the case. One possibility is that the addition of new assets independent of sufficient training in the deployment of those assets may result in inefficient and ineffective use. That is, the new assets may in fact be consumed at a higher rate than the previous assets (under the prior organization) because the task force commander is unable to incorporate them into the scheme of the maneuver. Whether this hypothesis is true or whether some other causes are at work can only be determined by the more comprehensive analyses to be performed in the future.

As indicated in the Data Analysis Section, the analysis of force-on-force performance data was necessarily limited by the type and level of information available for analysis. While recognizing those limitations, it was felt that there was much to be gained by exploring performance variations to determine the extent of their existence and whether they were related to the macro-level variables available in the Take Home Packages. The results of this analysis have accomplished that objective. They have demonstrated that variation does exist and that the macro-level variables while related do not account for all of the observed differences in performance. Further, the analyses have pointed to several areas to be pursued in the more comprehensive analyses to be conducted with the NTC digital data. Thus, the results reported here represent the jumping off point for this new research effort.

TITLE: Use of Instrumentation to Improve Combat Readiness
AUTHOR: William L. Shackelford
AFFILIATION: The BDM Corporation

1. Introduction: (Slide 1) Soldiers sit hunched in their fighting positions intently peering into the still desert darkness.

The task force intelligence officer is frantically attempting to contact his scouts by radio who are twenty kilometers to the front looking for the enemy. The commander is being briefed on the status of the defensive obstacle preparation within the task force sector. The brigade operations and intelligence net is crackling with sightings of a large enemy force on the move. Suddenly, the stillness is shattered by the deafening impact of steel from massed enemy artillery. The S-2's scream is barely audible that the scouts are in contact with a force of 150 to 190 armored vehicles. (Slide 2)

The enemy motorized rifle regiment is on the move and attacking into the guts of the task force defense.

This action is repeated over and over again at the Army's National Training Center. It is but one of the battles that U.S. Soldiers will fight during their two weeks of pre-combat hell that prepares them to meet the Warsaw Pact in the defense of Europe.

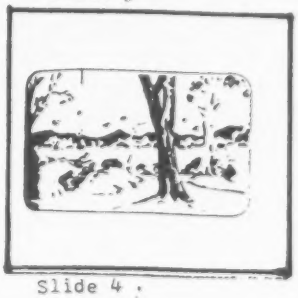
(Slide 3) The National Training Center was established in 1981 at Fort Irwin, California. Fort Irwin is located in the Mojave Desert at the point half way between Los Angeles and Las Vegas. The closest civilization is 40 miles away at the town of Barstow.

(Slide 4) The NTC provides a facility where units can conduct realistic training to develop individual, leader, and collective proficiency in combined arms operations. The environment replicates combat conditions more realistically than can be accomplished at home station due to physical limitations or resource constraints.

Slide 1

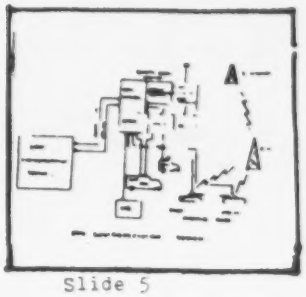
Slide 2

Slide 3


Slide 4

The NTC provides the Army with the ability to train and evaluate soldiers, leaders, and units in the execution of Army Airland Battle Doctrine. (Slide 5)

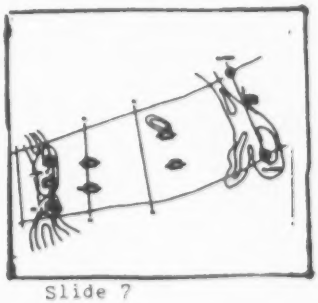
Unique to the NTC is an unobtrusive instrumentation system which captures the actions of the task force in realtime. Instrumentation is used as training feedback and as an instrument to provide insights and solutions to Army-wide issues of major significance. Battle engagements are recorded by the use of an integrated position location system and multiple integrated Laser Engagement System, called MILES, affixed to each combat vehicle. Opposing U.S. and enemy forces engage and cause casualties to each other through the use of eye-safe lasers which are attached to direct fire weapon system. Mobile field video units captire significant unit activities on tape. Command and control communications are recorded for later feedback to the unit. Exercise control and observer controller assessments are supported by the communications system. All field events are transmitted from the simulated battlefield through a remote range station to a central operations center. It is in the Core Instrumentation System that all information is fed into central computers and training analysis personnel assess the performance of the unit. (Slide 6)



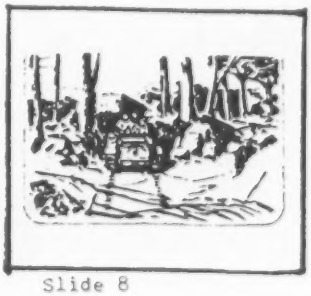
Slide 5

BLANK
Slide 6

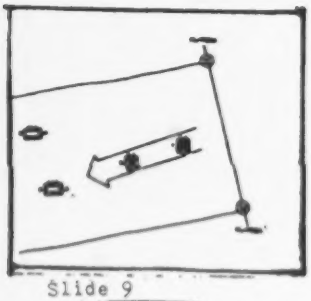
II. A Defense Scenario. The NTC instrumentation system is the most powerful training system available to any army. The instrumentation system allows the NTC trainers to gain insights into unit tactical performance not available through any other means. The results achieved by the instrumentation are fed back to the unit commander so that he may take the appropriate measures to correct his deficiencies. The following tactical scenario example will demonstrate the forceful utility of using instrumentation to improve the Army's fighting capability. (Slide 7)


Slide 7

At 0500 hours U.S. defending task force is deployed as follows: Scouts are to the front. A company has two platoons defending forward with one platoon and the company headquarters tanks to the rear. Behind them are B and C Companies and the task force headquarters tanks. (Slide 8)

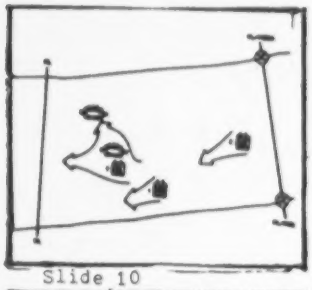

Slide 8

(Slide 9) At 0550 the motorized rifle regiment attacks into the task force defensive sector.



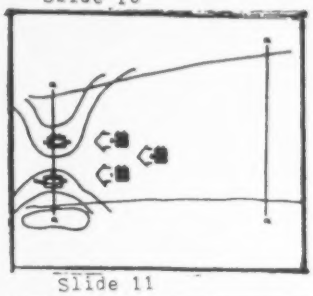
Slide 9

(Slide 10) At 0620 the lead motorized rifle battalion attacks the two forward defending platoons. With little loss of momentum the regiment reorganizes on the move after destroying the two forward platoons and drives deeper into the defenders position.

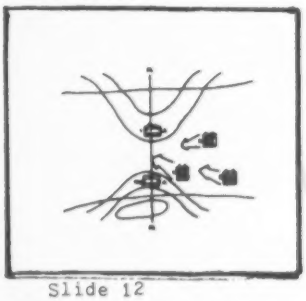


Slide 10

(Slide 11) At 0651 the regiment deploys to attack the remainder of A Company. At 0705 the regiment destroys the remaining A Company defenders who were deployed at the choke point in the terrain. (Slide 12)

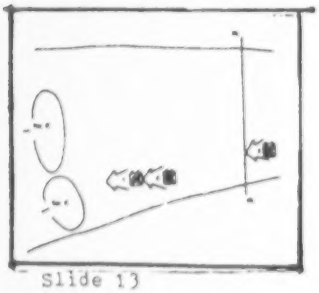


Slide 11



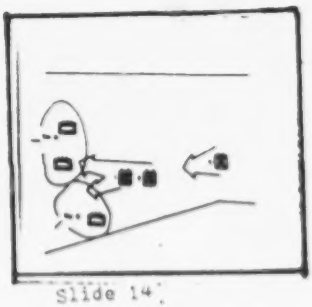
Slide 12

(Slide 13) At 0720 the regiment is deep into the defensive sectors 2000 meters from the defending B and C Companies.



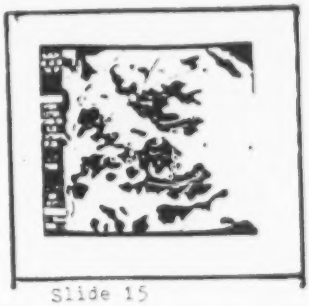
Slide 13

At 0731 the regiment attacks into the final defenses of the task force and at 0759 defense is broken with elements of the regiment driving toward their assigned objective of the day. (Slide 14)

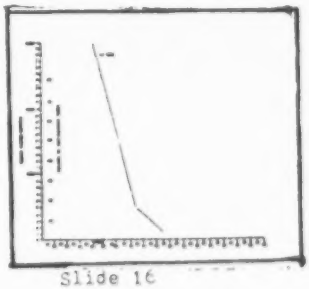

Slide 14

III. Analysis of Task Force Fire Control and Distribution of Tank Fires. What went wrong? The task force had 36 hours to prepare the defense. A lot of work has gone into preparing obstacles to slow the enemy. The defense plan was well understood by all the leaders. The commander knew this would be a tough fight but he never expected to be such an easy foe for the NTC opposing force regiment. These are some of the things that were going through the mind of the commander as he reflected on the battle. The NTC instrumentation system gives the NTC trainers the capabilities to point out serious problems to the commander so he can take the appropriate training corrective actions. The following example will demonstrate how a major problem was revealed in the fundamentals of tank fire control and distribution which led to the defeat of the task force by the regiment.

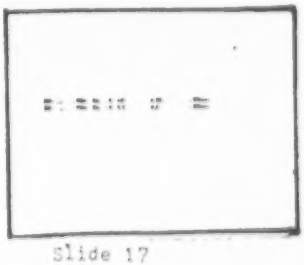
(Slide 15) The location and movement of the motorized rifle regiment in relation to the distance from the defenders was recorded using the color graphics position location capability of the system.


Slide 15

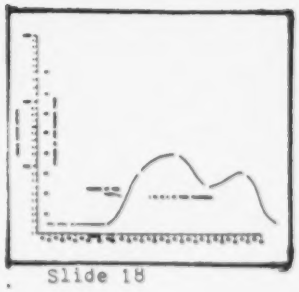
(Slide 16) The location of the opposing force is plotted on a graph by time and distance from the defenders.


Slide 16

(Slide 17) Tank main gun firing events are captured from the system data base to reflect the volume of fire by the defenders.


Slide 17

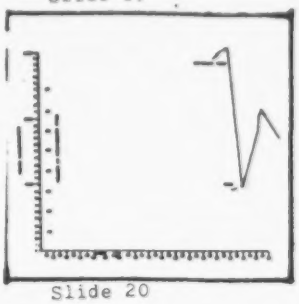
(Slide 18) These tank firing events are plotted on a graph reflecting the number of rounds fired by the units at selected time intervals.



Slide 18

(Slide 19) The final element of information needed is to plot the effectiveness of the tank firings against the attacking regiment. Hits, kills, and near misses are recovered from the data base and placed upon a graph by event, range, and time the hit, kill, or near miss occured. (Slide 20)

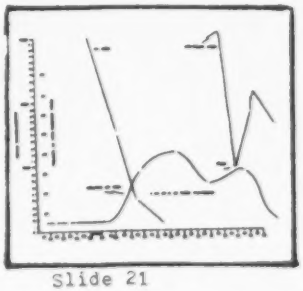
Slide 19



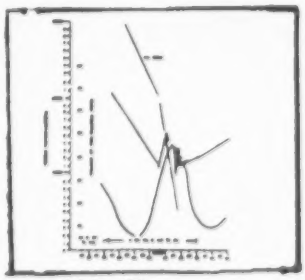
Slide 20

(Slide 21) By superimposing all three elements on a single graph a composite picture is then presented which depicts the effectiveness of the friendly unit's fire contol and distribution techniques. This composite pictures shows the initial fight of the forward platoons against the regiment. The enemy attack was unempeded and friendly platoons were overrun.

(Slide 22) The volume of fire never increased until the enemy had overrun the platoon. The friendly fires were not effective and only resulted in killing four enemy vehicles and only after the platoons were overrun. This slide shows the results of the encounter between the remaining A Company platoon and the Company Command element. Once again, the enemy is allowed to close in to a range of 1500 meters before the volume of fire and fire effectiveness causes eleven enemy vehicle kills. This is too little and too late because the motorized rifle battalion continues to press through the defenders.



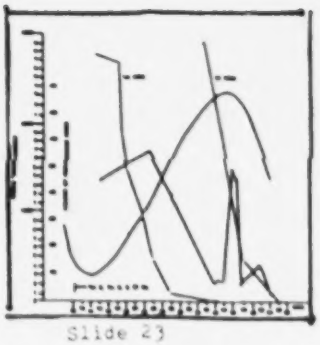
Slide 21



Slide 22

(Slide 23) The enemy regiment continues to press the attack into the last defenses of the task force. The 1st motorized rifle battalion is successful in closing on the defenders to be followed shortly by the 2nd motorized rifle battalion who completes the penetration. The friendly volume of fire is negligible against the 1st MRB and only increases when the defenders are being overrun by both MRB's.

Fire effectiveness is again negligible with only a total of six opposing force combat vehicles being killed by the defender's tanks. (Slide 24)



Slide 23

BLANK

Slide 24

IV. Summary. The use of NTC instrumentation as you have seen during this presentation has a major impact on increasing the combat readiness of the U.S. Forces. Armed with the clear picture of a serious problem he has within his task force, the commander can now take specific action to correct this deficiency. The NTC instrumentation system gives the commander the tools to improve and adjust homestation training so that he is prepared to accomplish his mission in combat. (Slide 25)

Slide 25

Collecting and Utilizing Feedback: Combined Arms and Services Staff School

Ralph W. Ekwall, Ed. D.
Ft. Leavenworth, Kansas

Since the Combined Arms and Services Staff School (CAS^3) is one of four schoolhouses within the Command and General Staff College (CGSC) one may ask why two different feedback systems exist within this college. The answer is related to the substantial differences in the two programs. Since the programs differ greatly, their feedback requirements are also different. Without going into great detail, they differ in these respects: length, rank of students, class size, curriculum design, evaluation system, and scope of instructor responsibility. Some of these differences will be discussed in detail in the following section.

Background Information

Some background information is needed to fully understand the feedback program. The Combined Arms and Services Staff School which we call CAS cubed is a fairly new school which was created to teach staff skills to Army captains. The curriculum is divided into two phases: Phase I and Phase II. Phase I is the non-resident portion which must be completed prior to attending the resident portion; it consists of 15 sub-courses called modules. The average student can complete Phase I in about 140 clock hours. Phase II is the nine-week resident portion of the course. Phase II consists of six exercises varying in length from 4 to 15 days. Phase II exercises are subdivided into lessons. Every student is assigned to a 12-person Staff Group. This Staff Group is taught by an experienced lieutenant colonel called a Staff Leader who is responsible for the entire block of Phase II instruction. Since the same Staff Group stays with one Staff Leader for nine weeks, strong bonds of friendship are developed. The final course grade is a pass-fail grade; the emphasis is on providing immeidate feedback on student performance for purposes of improvement. We also emphasize the development of higher level mental skills such as analysis, snythesis and problemsolving; these skills are emphasized more than the mastery of information. Hence, there are no exams in Phase II. This course represents the third mandatory educational experience for an Army captain; the first is the Officer Basic Course, the second is the Officer Advanced Course, and the third is CAS^3. This course may be the first course where the officer attends a school with officers of other branches. The CAS^3 program makes Staff Group assignments to achieve maximal mixing of branches. The CAS^3 course was extensively revised after its first pilot iteration in 1981. There was further revision after the second iteration. At present we are on a yearly rewrite schedule. Revision information is collected throughout the year and then used as a basis for a planned rewrite. REvision information consists of information such as the following: information extracted from doctrinal changes that must be integrated into the curriculum, command guidance, error correction and information collected from the CAS^3 Student Perception of Curriculum Relevance (SPCR) instruments. All of these data sources are utilized in annual curriculum revisions; in this presentation the concentration is on creation, administration and utilization of SPCR. The CAS^3 Coordinating Author's office has the responsibility of an after-action report to the Director after every nine-week class is completed. The SPCR instrument is the basis for this report.

Purpose

The SPCR instrument was developed for the following reasons. First of all, there was a need to collect student perceptions of the course so that curriculum writers and the Coordinating Author would know if the difficulty level was too high or too low. They needed to know if the course was perceived as relevant, if the amount of time spent on various topics was appropriate, and if students perceived the content as being current. This is a good place to digress briefly on the topic of perception. Student perceptions do not equate to student achievement. Student perceptions tend to be soft data rather than hard data. Nevertheless, student perceptions are real; if students perceive a course as being irrelevant, too easy, too hard, or poorly presented, then the perception may be sufficiently real to cause the course to be poorly received or it may cause the course to fail. Student perceptions are a fundamental basis for the success of any course. To get back to the original topic -- we knew that student perceptions of CAS^3 were a critical element and we needed an instrument to collect these perceptions for purposes of curriculum revision. A second reason for the development of the SPCR instrument was to provide feedback to Staff Leaders about the functions of their Staff Group and the morale of their Staff Group in comparison with other Staff Groups, and to do this on a timely basis so that corrective action could be taken. Another aspect of the feedback process was providing information to the Director about morale and progress of the various Staff Groups.

Explanation of the System

The CAS^3 program is 42 days long. The SPCR questionnaire is adminsitered three times. It is administered on day 8, on day 22, and day 37. The format is such that eight questions are asked about each lesson.

BEST COPY AVAILABLE

TABLE I

	Responses - A--Strongly Agree, B--Agree, C--Neutral, D--Disagree, E--Strongly Disagree	
	Lesson 1, Staff Techniques: Problemsolving	Lesson 2, Staff Techniques: Military Writing
This lesson was valuable for an officer in my branch.	1. _____	9. _____
Skills and information learned in this lesson will be useful to me in future assignments.	2. _____	10. _____
Feedback, critiques and after action reviews helped me to learn and improve my skills.	3. _____	11. _____
Phase I materials did a good job of preparing me for this lesson.	4. _____	12. _____
This lesson supported CAS^3 goals.	5. _____	13. _____
The material in this lesson is current.	6. _____	14. _____
	Responses - A--Too High, B--High, C--About Right, D--Low, E--Too Low	
Judge the amount of time allocated to this lesson.	7. _____	15. _____
Judge the difficulty level of this lesson.	8. _____	16. _____

We want to know if the lesson was relevant for an officer in that particular branch; we ask if knowledge and skills learned will be useful in future assignments; we ask about feedback; we ask about Phase I preparation; we ask if the lesson supports our four basic goals; we ask if lesson material is current; and we ask about difficulty and time allocations. Note that questions 1-8 deal with Lesson 1 and questions 9-16 ask the same 8 questions about Lesson 2 and so on. Students are asked to give responses ranging from A=strongly agree to E=strongly disagree for each question. On questions 7 and 8, A=too high and E=too low. Students write their responses on an IBM answer sheet. Staff Leaders have the option of requiring students to first write their response on the questionnaire before putting responses on the IBM sheet. The advantage of this procedure is that the Staff Leader receives feedback about his Staff Group's responses on an immediate basis. Questions are asked about lessons and also about exercises.

BEST COPY AVAILABLE

TABLE 2

	Responses - A--Strongly Agree, B--Agree, C--Neutral, D--Disagree E--Strongly Disagree
	Staff Techniques Exercise
Morale in the Staff Group was good.	89. _____
Everyone in the Staff Group did their share of work.	90. _____
	Responses - A--Very High, B--High, C--About Right, D--Low, E--Very Low
Judge the stress level.	91. _____
Judge the workload.	92. _____

The last four questions refer to an entire exercise. They ask about morale, Staff Group functioning, stress level, and workload. The response forms are processed and the response data is placed in an Statistical Package for the Social Sciences (SPSS) file for analysis. This file contains information for the CAS^3 data base on branch affiliation and Staff Group membership. The branch and Staff Group data provide us with a powerful tool for analysis of this data. Examples of analysis will be discussed in detail in the section following.

Examples of Use

Army branches can be classified into four groups. (1) Combat Arms such as Infantry and Armor, (2) Combat Support such as Military Police, (3) Combat Service Support such as Finance and Adjutant General and (4) Professionals such as Army Nurse and Judge Advocate General. The professional branches usually receive professional training prior to entering the Army. They usually receive direct commissions and they usually enter the Army at a rank higher han second lieutenant. Their work in the Army resembles the work of their civilian counterparts more than it resembles the work of other officers in the Army. Our program wanted to know how these professionals perceived our program. We already know that the overall perception was positive, but we were not sure if this perception was shared by all branch groups. Nor did we know if different parts of our program were perceived equally by different branch groups. At this point a reminder is necessary. There are six exercises in the Phase II resident portion. The first exercise called Staff Techniques teaches students how to use several quantitative techniques. It teaches them how to prepare a variety of Army documents. The remainder of the exercises deal with Training Management, Budget, Mobilization, Preparation for Combat and fighting a war in Europe. First we looked at the professionals and the other branch groups in reaction to question 9 on the first administration:

TABLE 3

Lesson 2, F121, Military Writing
(First Administration)

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Combat Arms	79.8%	14.9%	5.3%	--	--
Combat Support	83.7%	11.6%	4.7%	--	--
Combat Service Support	75.0%	15.6%	9.4%	--	--
[illegible]	[illegible]	[illegible]	1	[illegible]	--

1038

BEST COPY AVAILABLE

"This lesson was valuable to an officer in my branch." This lesson topic was Military Writing. One thing is remarkable: there are no disagrees or strongly disagrees. Note that there is less enthusiasm among the professionals, but they apparently feel the lesson is valuable even though they show less enthusiasm.

TABLE 4

Lesson 5, F121, Military Briefing
(First Administration)

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Combat Arms	84.0%	13.8%	2.1%	--	--
Combat Support	81.4%	16.3%	2.3%	--	--
Combat Service Support	81.3%	18.8%	0%	--	--
Professional	66.7%	26.7%	6.7%	--	--

Next, lets look at question 73 which refers to Lesson 5, Military Briefing and asks if that lesson was valuable to an officer in that branch. The response here is similar -- just less enthusiasm. Next, lets look at question 97, third administration of the SPCR.

TABLE 5

Lesson 12, F626, Division Operations Plan
(Third Administration)

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Combat Arms	42.9%	45.7%	9.5%	1.9%	0%
Combat Support	27.7%	66.0%	6.4%	0%	0%
Combat Service Support	25.7%	51.4%	20.0%	.9%	0%
Professional	26.3%	42.1%	26.3%	0%	5.3%

The item asks how valuable the lesson on the Division Operation Plan is to an officer in that branch. This data shows that the perception of professional branch officers is similar to Combat Service Support and Combat Support branch groups. Our conclusion was that the professional branch members perceived some lessons less positively, but no action was required on this matter. We will now consider another example of use of the SPCR. Remember that students are required to complete a set of 15 modules by correspondence prior to attending the resident portion. We wanted to find out if students felt well prepared for the Phase II lessons. We looked at responses to the question, "Phase I material did a good job of preparing me for this lesson."

BEST COPY AVAILABLE

TABLE 6

Phase II Lessons for Which Students
Felt Poorly Prepared

Lesson	Percentage Who Agree and Strongly Agree
Lesson 6, F121, Meeting Management	36%
Lesson 4, F121, Time Management	37%
Lesson 3, F121, Quantitative Skills	64%

We found two lessons where students felt poorly prepared. The lessons were Lesson 2; Time Management and Lesson 4, Meeting Management both of which were in the Staff Techniques exercise. The responses to Lesson 3, Quantitative Skills are included for purposes of comparison. It is the intent of our program to use Phase I to prepare students for Phase II, but in this instance we identified an oversight. This oversight is being considered by the curriculum committee.

The SPCR was used to analyze student perceptions of the Staff Ride. The Staff Ride is used to teach students how to analyze terrain. Students are taken by bus to various sites to practice terrain analysis uder the direction of Staff Leaders. There were some complaints from students so we decided to do an analysis of this lesson.

TABLE 7

Lesson 4, F626, The Staff Ride
(Third Administration)

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Combat Arms	31.4%	41.0%	18.1%	7.6%	1.9%
Combat Support	36.2%	38.3%	14.9%	10.6%	0%
Combat Service Support	20.0%	34.3%	31.4%	8.6%	5.7%
Professional	36.8%	36.8%	21.1%	5.3%	0%
Total	31.0%	38.6%	20.0%	8.1%	2.4%

When we looked at it, this is what we saw. One analysis technique we often use is to combine agree and strongly agree responses. Here, the combined figures add up to 69% which is a more than 2/3 approval.

TABLE 8

Lesson 7, F626, The Personnel Estimate
(Third Administration)

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Combat Arms	28.6%	41.9%	25.7%	2.9%	
Combat Support	14.9%	48.9%	29.8%	4.3%	
Combat Service Support	34.3%	54.3%	8.6%	2.9%	
Professional	26.3%	57.9%	10.5%	0%	
Total	26.2	[illegible]	21.9	2.9	

When we compared it to another, more or less representative lesson on the Personnel Estimate, the approval ratings were similar. In this case we decided to do a little more analysis and to look at responses according to Staff Leaders. We looked at responses to the question, "Skills and information learned in this lesson will be useful to me in future assignments."

TABLE 9

Responses Analyzed According to Staff Leaders

Lesson 4, F626, The Staff Ride

Staff Leader Identifier	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
D	3	4	3	--	--
G	1	8	2	1	--
B	-	6	3	1	3
F	3	2	3	3	2

(Analysis by Numbers)

Here are two examples of staff leaders who apparently convinced their students that the Staff Ride was critical to their future success and two other Staff Leaders who were not so successful. We decided, as a result of this analysis, to make sure that incoming Staff Leaders received training on how to conduct a Staff Ride. This training has been implemented and we now hope to see an improvement in student perception of the Staff Ride.

The SPCR instrument is used in an analytical mode, but it is also used to provide feedback to instructors. You may remember that questions were asked about exercises in addition to the 8 questions asked bout each lesson. Students are asked about their perception of morale in the Staff Group; they are also asked if everyone in the Staff Group is doing their share of work. Here is an example of the kinds of feedback provided to the Staff Leader. This is a response to the question, "Morale in the Staff Group is good."

TABLE 10

Example of Feedback Provided to Staff Leaders

Staff Leader Identifier	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
A	4	6	0	1	0
H	0	1	9	0	0
P	6	6	0	0	0
O	2	3	3	3	0

(Analysis by Numbers)

Here is an example. Each Staff Leader receives information on all of the Staff Groups; he will know that his code letter is F or N or whatever, but he will not know the code letter of other Staff Leaders. This gives a Staff Leader a chance to compare morale in his Staff Group with other Staff Groups. Since there are three administrations of this instrument, the Staff Leader can look at the status of morale at three different time periods. The Staff Leader can also compare morale levels with previous Staff Groups he has had. Note that in this instance Staff Leader H has nine people who are neutral about morale. Morale in group P appears to be good. The same kind of information is provided in regard to student perception of other Staff Group members doing their share of work. This information may cause that particular staff leader to re-look at a problem, or it may cause him to take corrective action.

SUMMARY

This system which asks the same 8 questions about each lesson and then analyzes the responses by Staff Group and branch membership has proved to be quite successful.

BY-LAWS OF THE MILITARY TESTING ASSOCIATION

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel and assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems.

3. Each agency shall have no more than two (2) representatives.

1042

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. Representation from the majority of the organizations of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to issue notices for conferences. The Secretary shall solicit nominations for the Harry Greer award prior to the annual conference. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The coordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

C. The annual conference of the Association shall be held at a time and place determined by the coordinating agency. The membership of the Association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

1043

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the Association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the Association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Harry H. Greer Award

A. Selection Procedures:

1. Recipients of the Harry H. Greer Award will be selected by a committee drawn from the agencies represented on the MTA Steering Committee. The CO of each agency will designate one person from that agency to serve on the Awards Committee. Each committee member will have attended at least three previous MTA meetings. The member from the coordinating agency will serve as chairman of the committee.

2. Nominations for the award in a given year will be submitted in writing to the Awards Committee Chairman by 1 January of that year.

3. The Chairman of the committee is responsible for canvassing the other committee members to arrive at consensus on the selection of a recipient of the award.

4. No more than one person is to receive the award each year, but the award need not be made each year. The Awards Committee may decide not to select a recipient in any given year.

5. The annual selection of the person to receive the award or the decision not to make an award that year is to be made at least six weeks prior to the date of the annual MTA Conference.

BEST COPY AVAILABLE

B. Selection Criteria:

1. The recipients of the Harry H. Greer Award are to be selected on the basis of outstanding work contributing significantly to the MTA.

C. The Award:

1. The Harry H. Greer Award is to be a certificate normally presented to the recipient during the Annual MTA Conference. The awards committee is responsible for preparing the text of the certificate. The coordinating agency is responsible for printing and awarding the certificate.

Article XII - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 21 October 1985).

MTA STEERING COMMITTEE MEMBERS

US Army Research Institute

US Air Force Human Resources Laboratory

US Air Force Occupational Measurement Center

US Coast Guard Institute

US Naval Education and Training Program Development Center

US Navy Personnel Research and Development Center

Belgian Armed Forces Psychological Research Section

Canadian Forces Directorate of Military Occupational Structures

Canadian Forces Personnel Applied Research Unit

Federal Republic of Germany Ministry of Defense

Royal Australian Air Force Evaluation Division

National Headquarters Selective Service System

US Navy Occupational Data Analysis Center

Defense Activity for Non-Traditional Education Support

1045

MINUTES

MTA STEERING COMMITTEE MEETING
27TH ANNUAL CONFERENCE
21-25 OCTOBER 1985
SAN DIEGO, CALIFORNIA

OPENING REMARKS

The meeting was opened by Dr. Martin Wiskoff, the Conference Chairman for 1985.

A list of the Steering Committee members present at this meeting is attached.

The minutes of the 26th MTA Conference held in Munich, Germany were reviewed and unanimously approved by the Steering Committee.

REQUEST FOR MEMBERSHIP

The Navy Occupational Data Analysis Center and the Defense Activity for Non-Traditional Education Support requested Primary Membership and were approved.

HARRY GREER AWARD

A written nomination was submitted in December 1984 by COL Paul T. Ringenbach, Commander, USAF Occupational Measurement Center for Mr. Fred Hawrysh, of the Directorate of Military Occupational Structures, National Defense Headquarters, Ottawa, Canada for the 1985 Harry H. Greer Award. Following discussion, the Steering Committee voted in favor of the nomination.

BY-LAW CHANGES

Dr. John Ellis of the Navy Personnel Research and Development Center proposed that the by-laws be amended to include greater mention of training research and development. After some discussion it was decided not to change the by-laws, but to provide greater emphasis to training in the annual call for papers and to expand the mailing list to include a broader range of training organizations.

Two changes to the by-laws were approved. The intent of Article V, Section A of the by-laws was clarified to read "3. Each agency shall have no more than two representatives." A listing of current members of the MTA Steering Committee is to be included at the end of the by-laws. The revised by-laws appear in the 1985 proceedings.

PROCEEDINGS GUIDELINES

There was discussion as to the desirability of establishing guidelines for submitting papers. A proposal will be made by the Federal Republic of Germany.

MTA 1986

The U. S. Coast Guard Institute is scheduled to host the 1986 meeting, either in Oklahoma City, OK or in the New London, CT area. The 1986 MTA President will be CAPT Robert W. Davis and the Chairman will be Mr. John Burt.

MTA 1987-1988

The Canadian Forces Directorate for Military Occupational Structures and the Canadian Forces Applied Research Unit will host the 1987 meeting in Ottawa and the Army Research Institute will host the 1988 meeting in an undetermined location.

ADJOURNMENT

There being no further business, the meeting was adjourned.

STEERING COMMITTEE MEMBERS PRESENT

21 October 1985

Carol V. Ascherfeld	U.S. Navy Education and Training Program Development Center
Walter W. Birdsall	U.S. Navy Education and Training Program Development Center
Arnold Bohrer	Belgian Armed Forces, Psychological Research Section
John A. Burt	U.S. Coast Guard Institute
John Ellis	U.S. Navy Personnel Research and Development Center
SQNLDR Ken Given	Royal Australian Air Force Evaluation Division
Roger G. Goldberg	U.S. Defense Activity for Non-Traditional Education Support
Wulf Gronwald	Federal Republic of Germany, Ministry of Defense
Robert F. Holz	U.S. Army Research Institute
CDR J. E. Olson	U.S. Navy Education and Training Program Development Center
LCOL Terry J. Prociuk	Canadian Forces, Personnel Applied Research Unit
Martin L. Rauch	Federal Republic of Germany, Ministry of Defense
COL P. T. Ringenbach	U.S. Air Force Occupational Measurement Center
J. S. Tartell	U.S. Air Force Occupational Measurement Center
Raymond O. Waldkoetter	U.S. Army Research Institute
Robert J. Wilson	U.S. Navy Occupational Data Analysis Center
Martin Wiskoff	U.S. Navy Personnel Research and Development Center
COL G. A. Zypchen	Canadian Forces, Directorate of Military Occupational Structures

Military Testing Association
1985
Harry H. Greer Award
to
FRED HAWRYSH

The Harry H. Greer Award is hereby presented to Mr. Fred Hawrysh of the Directorate of Military Occupational Structures, National Defence Headquarters, Ottawa, Canada, for outstanding work contributing significantly to the goals and purposes of the Military Testing Association.

You have been an active proponent of MTA for many years, first as an active-duty representative of the Canadian Forces and later as the senior civilian for your agency. We acknowledge your encouragement of MTA attendance and your leadership in stimulating quality presentations.

Your achievements in stimulating international exchange and cooperative work in the field of occupational analysis are outstanding. The continuing success of the Comprehensive Occupational Data Analysis Programs (CODAP) reflects positively on your dedication and enthusiasm.

The devotion and professionalism of your work are of great credit to yourself and the association. With gratitude, appreciation, and friendship of all those affiliated, the MTA recognizes Fred Hawrysh with this Harry H. Greer Award.

MTA
THE MILITARY TESTING ASSOCIATION

Howard S. Eldredge
MTA President
Commanding Officer
Navy Personnel Research
and Development Center

MTA 27TH ANNUAL CONFERENCE STAFF

MTA PRESIDENT

CAPTAIN HOWARD S. ELDREDGE

MTA CHAIRMAN

DR. MARTIN F. WISKOFF

COMMITTEE CHAIRPERSONS/MEMBERS

PROGRAM COMMITTEE

DR. JOHN PASS
DR. JOHN ELLIS
DR. NORM ABRAHAMS

OPERATIONS COMMITTEE

MR. ROB WELLS
MRS. MARGIE SANDS
MS. PAT MARSH
MR. JIM JULIUS
MS. JESSIE GRIER
ET1 JIM LAW

FINANCE

MS. SUE STUMPF

PUBLICATIONS/GRAPHICS

MR. GENE STOUT
MS. DARLA LEITHISER
MR. PORTER WOOTEN
MR. TED YELLEN
MS. CARMEN SCHEIFERS

HOSPITALITY

DR. JOHN ELLIS
DR. JOHN PASS

"A"

*DR. NORM ABRAHAMS
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*HOMER W. ADKINS
US NAVY CHIEF OF NAVY EDUCATION
AND TRAINING (CNET N-131)
NAS PENSACOLA, BLDG. #624
NAS PENSACOLA FL 32508

*DR. JEFFREY W. ANDERSON
12760 GAZEBO COURT
WOODBRIDGE, VA 22192

*WAYNE A. ANDERSON
HQ USMEPCOM MEPCT-M
2500 N. GREENBAY RD.
NORTH CHICAGO, IL 60064-3094

*ROBERT J. ANGUS
CANADIAN FORCES PERSONNEL APPLIED
RESEARCH UNIT
4900 YONGE STREET, SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

*MR. THOMAS M. ANSBRO
CNET N-2XX CNET HQ
NAS
PENSACOLA FL 32508

*DR. JANE M. ARABIAN
COMMANDER
U.S. ARMY RESEARCH INSTITUTE
PERI-RS (ATTN: Dr. ARABIAN)
5001 EISENHOWER AVE.
ALEXANDREA, VA 22333-5600

*DR. CAROL V. ASCHERFELD
NAVEDTRAPRODEVCEN (CODE 3143)
SAUFLEY FIELD
PENSACOLA, FL 32509-5000

*LEANNE E. ATWATER
NAVY PERSONNEL R&D CENTER
CODE 04
SAN DIEGO, CA 92152-6800

*NANCY K. ATWOOD
ARI, PRESIDIO OF MONTEREY FIELD UNIT
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

"B"

*MS. ANNETTE G. BAISDEN
NAVAL AEROSPACE MEDICAL RESEARCH INST.
NAMI (CODE 11)
PENSACOLA FL 32505

*DR. HERBERT GEORGE BAKER
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*MR. THOMAS B. BAKER
COMMANDER
U.S. ARMY ORDNANCE CENTER & SCHOOL
ATTN: ATSL-TD-IUA (MR. BAKER)
ABERDEEN PROVING GROUND, MD 21005-5201

*DR. JAMES H. BANKS
U.S. ARMY RESEARCH INSTITUTE
FIELD UNIT
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*WILLIAM G. BARNES
COMMANDANT
U.S. ARMY CHEMICAL SCHOOL
ATTN: ATZN-CM (MR. BARNES)
FT. MCCLELLAN, AL 36205-5000

*MS.JOANN BARSIS
PERSONNEL PERFORMANCE INC.
3304 BUCKEYE LANE
FAIRFAX, VA 22033

*DR. WILLIAM M. BART
EDUC PSYCHOLOGY
330 BURTON HALL
UNIVERSITY OF MINNESOTA
178 PILLSBURG DRIVE S.E.
MINNEAPOLIS, MN 55455

*HELEN E. BELLETTI
1908 CHRIS SCOTT DRIVE
EL PASO, TX 79936

*STEPHEN G. BENIGNI
NODAC
BLDG 150, WASH. NAVY YARD
WASHINGTON, D.C. 20374

*CAPT JEANMARIE BESSET
AIR FRANCE DONJ
ADMINISTRATIVE
BP 10201
95703 ROISSY CH DEGAULLE CEDEX
FRANCE

*MAJ CONRAD G. BILLS
INSTRUCTIONAL SYSTEMS DEV DIV
93 BMW/DOS
CASTLE AFB, CA 95342-5000

*WALTER W. BIRDSALL
NAVEDTRAPRODEVCEN (CODE 31)
SAUFLEY FIELD
PENSACOLA FL 32561

*CAPT JACK L. BLACKHURST
AFHRL/MODP
BROOKS AFB, TX 78235-5000

*HARRIET H. BLAKELY
US MEPCOM
MILITARY ENTRANCE PROCESSING ST.
(ATLANTA)
80 TENTH ST. N.E.
ATLANTA, GA 30309-3957

*BRUCE BLOXOM
NAVAL POSTGRADUATE SCHOOL
P.O. BOX 222375
CARMEL, CA 93922

*FREDERICK L. BLUME
CNET
3595 BROOKSHIRE DRIVE
PENSACOLA, FL 32504

*DENNIS M. BYLTH
ARMY PERSONNEL RESEARCH ESTABLISHMENT
C/O RAE FARNBOROUGH
FARNBOROUGH
HAMPSHIRE, UNITED KINGDOM

*ARNOLD C. BOHRER
SELECTIE-EN REKRUTERINGS CENTRUM
SECTIE PSYCHOLOGISCH ONDERZOEK
BRUYNSTRATT
B-1120 BRUSSEL, BELGIUM

*JOHN BONDARUK
NATIONAL SECURITY AGENCY
FORT GEORGE G. MEADE
MARYLAND, 20755-6000
ATTN: M324

*RICHARD L. BRANCH
HQ US MEPCOM/MEPC-T
2500 GREENBAY RD.
NORTH CHICAGO, IL 60060

*JOHN S. BRAND
CDR USASSC
ATTN: AISG-ES (BRAND)
FORT BENJAMIN HARRISON IN 46216-5370

*HENRY I. BRAUN
ETS (22T) - RESEARCH
PRINCETON, NJ 08541

*ROBERT C. BREEST
COMMANDER, NAVY RECRUITING AREA FIVE
NTC GREAT LAKES, BLDG 3
GREAT LAKES, IL 60088

*JOEANNE BRIDGE
ALBUQUERQUE MEPS
P.O. BOX 103
ALBUQUERQUE, NM 87102-0103

*DR. BRENT BRIDGEMAN
EDUCATIONAL TESTING SERVICE
11-R, ROSEDALE ROAD
PRINCETON, NJ 08541

*ARTHUR M. BROWN
US MEPCOM
HQ E SECTOR
FORT MEADE, MD 20755

*DR. JOHN R. BRUNI
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*JANICE BUCHHORN
AFHRL/TSOZM
BROOKS AFB TX 78235

*COMMANDING OFFICER
ATTN: LINDA BURGUM
VS41, BLDG. 797
NAS, NORTH ISLAND
SAN DIEGO, CA 92135

*LAWRENCE F. BURNS
NAVY RECRUITING COMMAND
4015 WILSON BLVD.
ARLINGTON, VA 22203

*JOHN A. BURT
USCG INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

*LLOYD D. BURTCH
AFHRL/XR
BROOKS AFB TX 78235

*BRIAN J. BUSH
ARI
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*WALTER G. BUTLER
DIRECTOR, US ARMY TRADOC SYSTEMS
ANALYSIS ACTIVITY
ATTN: ATOR-THD
WHITE SANDS MSL RNG, NM 88002-5502

"C"

*WAYNE J. CAMARA
HUMAN RESOURCES RESEARCH ORG.
1100 SOUTH WASHINGTON ST.
ALEXANDRIA, VA 22314

*EDMUND J. CARBERRY
US ARMY ARMOR SCHOOL
7303 RIDAN WAY
LOUISVILLE, KY 40214

*CW03 PETER F. CASE
U.S. COAST GUARD (PMR-5)
2100-2ND STREET S.W.
ROOM 4406
WASHINGTON, D.C. 20593-0001

*JEANNA F. CELESTE
WESTAT, INC.
1650 RESEARCH BLVD.
ROCKVILLE, MD 20854

*DR. FREDERICK R. CHANG
NAVPERSRANDCEN
CODE 51
SAN DIEGO, CA 92152-6800

*ROBERT E. CHATFIELD
NAVPERSRANDCEN
CODE 623
SAN DIEGO, CA 92152-6800

*JOAN T. CHIPPENDALE
US ARMY RESEARCH INSTITUTE
PERI-IR
BLDG. 501
FORT RUCKER, AL 36362

*GAYLE C. CHRISTIANSEN
US COAST GUARD HQ
2100 2ND ST SW
WASHINGTON, DC 20593

*DR. BARBARA R. CHURCH
BLDG. 262 RM. T-2
FEDERAL LAW ENFORCEMENT TRA CTR
GLYNCO, GA 31524

*RAY W. CLARK
NAVAL MILITARY PERSONNEL COMMAND DET
NAVY OCCUPATIONAL DEVELOPMENT AND
ANALYSIS CENTER
WASHINGTON NAVY YARD (ANACOSTIA)
BLDG 150
WASHINGTON, D.C. 20374-1501

*ANNE COATES-RADLOFF
HUMRRO
1100 SO. WASHINGTON STREET
ALEXANDRIA, VA 22314

*LCDR PAUL M. CONNOLLY
CNET CODE N-3324
NAS PENSACOLA, FL 32508

*MAJ PAUL COOK
HQ AFMPC/DPMYOT
RANDOLPH AFB, TX 78150-6001

*MERRI-ANN COOPER
ADVANCED RESEARCH RESOURCES
ORGANIZATION
4330 EAST-WEST HIGHWAY
BETHESDA, MD 20814

*DR CHARLES H. CORY
NAVY PERSONNEL R&D CENTER
CODE 621
SAN DIEGO CA 92152-6800

*DR. KENT CRAWFORD
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

*NEAL R. CROWLEY
5111 SOUTH 8TH ROAD #101
ARLINGTON, VA 22204

"D"

*MAJ MICKEY R. DANSBY
LMDC/AN
BLDG 678
MAXWELL AFB, AL 36112-5712

*DR. CHARLES E. DAVIS
OFFICE OF NAVAL RESEARCH
800 NORTH QUINCY STREET, CODE 442
ARLINGTON, VA 22217

*GENEVA DAVIS
3617 DALE HOLLOW RD.
ANNISTON, AL 36201

*RENEE DAVIS
6940 FORUM STREET
SAN DIEGO, CA 92111

*LEE ELLYN DAY
REHAB GROUP INC.
1360 ROSECRANS, SUITE D
SAN DIEGO, CA 92106

*DR. ESTHER E. DIAMOND
EDUCATIONAL & PSYCHOLOGICAL
CONSULTANT
721 BROWN AVENUE
EVANSTON, IL 60202

*CAPT BEN L. DILLA
AFIT/LSB
WRIGHT-PATTERSON AFB, OH 45433

*RONNA F. DILLON
DEPT. OF GUID. & EDUC. PSYCH.
SOUTHERN ILLINOIS UNIVERSITY
CARBONDALE, IL 62901

*DR. STEVE DOCKSTADER
NAVY PERSONNEL R&D CENTER
CODE 72
SAN DIEGO CA 92152-6800

*DR. LINDA M. DOHERTY
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO CA 92152-6800

*WILLIAM J. DOHERTY
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

*RICHARD D. DOORLEY
USA MILPERCEN (DAPC-MSE)
200 STOVALL STREET
ALEXANDRIA, VA 22332

*DR. WALTER E. DRISKILL
6422 FALLS CHURCH
SAN ANTONIO, TX 78247

*EUGENE H. DRUCKER
HUMRRO - FT. KNOX OFFICE
PO BOX 293
FORT KNOX, KY 40121

*CAPT. R. ERIC DUNCAN
AFIT/CIRP OL
UNIVERSITY OF TEXAS AT AUSTIN
1612 PLATEAU RIDGE
CEDAR PARK, TX 78613

"E"

*HEINZ-JURGEN EBENRETT
GERMAN ARMED FORCES OFFICE
(STREITKRAEFTEAMT)
POSTFACH 20 50 03
D-5300 BONN 2
FEDERAL REPUBLIC OF GERMANY

*JOHN M. EDDINS
CERL, UNIV OF ILLINOIS
103 S. MATHEWS ST., RM. 252
URBANA, IL 61801

*WILLIAM A. EDMUNDSON
ARMY EDUCATION CENTER
MACOMB & ARMISTREAD STREETS
FORT BRAGG, NC 28307-5000

*LINDA J. EISMAN
302 DRAKE ST.
BENICIA, CA 94510

*ANA G. EKSTROM
ARI FIELD UNIT
P.O. BOX 223133
CARMEL, CA 93922

*DR. RALPH W. EKWALL
CAS3-USACGSC
ATTN: ATZL-SWB (DR. RALPH EKWALL)
FORT LEAVENWORTH, KS 66027

*CAPT HOWARD S. ELDREDGE
NAVY PERSONNEL R&D CENTER
CODE 00
SAN DIEGO, CA 92152-6800

*DR. JOHN A. ELLIS
NAVY PERSONNEL R&D CENTER
CODE 51
SAN DIEGO CA 92152-6800

*TATIANA S. EROHINA
830 C AVE., APT. F
CORONADO, CA 92118

"F"

*BENJAMIN A. FAIRBANK
C/O PERFORMANCE METRICS, INC
5825 CALLAGHAN ROAD, SUITE 225
SAN ANTONIO, TX 78228

*FIORENZO FERRARI
ITALIAN AIR STAFF
STATO MAGGIORE AERONAUTICA
VIE UNIVERSITA ROMA
ITALY

*ROBERT L. FEUGE
MANTECH MATHETICS CORP.
P.O. BOX 26655
SAN DIEGO, CA 92126

*BERNARD J. FINE
HEALTH & PERFORMANCE DIVISION
USARIEM
ARMY R & D CENTER
NATICK, MA 01760

*DR. GERALD P. FISHER
SENIOR SCIENTIST
HUMRRO
1100 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

*ELI FLYER
HUMRRO
10 SIERRA VISTA DRIVE
MONTEREY, CA 93940

*THOMAS K. FORSYTHE
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

*RONALD E. FREEMAN
LOS ANGELES MEPS
5051 RODEO ROAD
LOS ANGELES, CA 90016

*ROBERT L. FREY, JR.
HEADQUARTERS, US COAST GUARD
G-P-1/2, ROOM 4200 B
2100 2ND STREET, S.W.
WASHINGTON, DC 20593

"G"

*DENNIS GAYNOR
TESTING DIRECTORATE
MEPCOM
1140 INVERRARY LANE
DEERFIELD, IL 60015

*LT.COL FRANK C. GENTNER, USAF
DEOMI/DSE
PATRICK AFB, FL 32925

*CHARLES M. GIBBONS
U.S. ARMY ORDNANCE MISSILE SCHOOL
ATSK-TIL
REDSTONE ARSENAL, AL 35897-6600

*DR. KAROL GIRDLER
U.S. ARMY RESERACH INSTITUTE
HQ AMC EUROPE
BOX 228
APO NY 09333

*KENNETH C. GIVEN
AIR FORCE HUMAN RESOURCES LAB
AFHRL/MODJ
BROOKS AFB
SAN ANTONIO, TX 78235

*DR. DWIGHT J. GOEHRING
ARMY RESEARCH INSTITUTE
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*ROGER G. GOLDBERG
DEPARTMENT OF THE NAVY, DANTES
CODE 10
PENSACOLA, FL 32509

*DR. LAWRENCE A. GOLDMAN
US ARMY SOLDIER SUPPORT CENTER-NCR
ATTN: ATZI-NOS-C (DR. GOLDMAN)
200 STOVALL STREET
ALEXANDRIA, VA 22332-0400

*DR. ALEXANDER M. GOTTESMAN
CHIEF, CURRICULUM SUPPORT DIVISION
CODE 221
NAVAL HEALTH SCIENCES EDUCATION
AND TRAINING COMMAND
NAVAL MEDICAL COMMAND,
NAT'L CAPITOL REG.
BETHESDA MD 20814-5022

*LINDA J. GRAHAM
HELICOPTER ANTI-SUBMARINE SQUADRON 10
TRAINING (ATTN: EDUCATION SPECIALIST)
BUILDING 626
NORTH ISLAND NAVAL AIR STATION
SAN DIEGO, CA 92103

*SCOTT E. GRAHAM
U.S. ARMY RESEARCH INSTITUTE
9109 POMONA PL.
LOUISVILLE, KY 40272

*MR.GERRY I. GREWELL
MILITARY ENTRANCE PROCESSING STATION
1755 FIFTH AVE.
SAN DIEGO, CA

*LAURA W. GRIEGER
COMMANDANT, USAIS
ATTN: ATSH-I-V-C-O (MS. GRIEGER)
FORT BENNING GA 31905-5593

*CAROL A. GRIFFITHS
TESTING SECTION, ROOM 1708
FEDERAL BUILDING
1000 LIBERTY AVENUE
PITTSBURGH, PA 15222-4101

*WULF GRONWALD
GAF INSTITUTE OF AVIATION MEDICINE
DEPT VI AVIATION PSYCHOLOGY
POSTFACH 1264/KFL
8080 FURSTENFELDBRUCK
FEDERAL REPUBLIC OF GERMANY
.kp off

"H"

*LCDR JANE P. K. HAMMOND
NAVGMSCOL, CODE 011
DAM NECK
VIRGINIA BEACH, VA 23462

*JAMES P. HANLON
230 E. QUEUR STREET
CHAMBERSBURG, PA 17201

*LAWRENCE M. HANSER
U.S. ARMY RESEARCH INSTITUTE
ATTN: PERI-RS (DR LAWRENCE M. HANSER)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*FRANK D. HARDING
ADVANCED RESEARCH RESOURCES ORG.
4330 EAST-WEST HIGHWAY
BETHESDA, MD 20814

*2ND LT DAVID L. HARDY
USAF OMC/OMYA
RANDOLPH AFB, TX 78150-5000

*GAIL R. HARDY
ROOM 423, SENIOR PSYCHOLOGIST (NAVAL)
ADMIRALTY ARCH IV
SPRING GARDENS, LONDON
SW1A2BL, ENGLAND

*DR. JOAN HARMAN
US ARMY RESEARCH INSTITUTE
ATTN: PERI-IC
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*W. ANDREWW HARRELL, PH.D.
DEPARTMENT OF SOCIOLOGY
THE UNIVERSITY OF ALBERTA
TURY 5-21
EDMONTON, ALBERTA T6G 2H4
CANADA

*JANE HATCH
NATIONAL COMPUTER SYSTEMS
1200 NEW HAMPSHIRE AVE. N.W.
SUITE 360
WASHINGTON, D.C. 20036

*FREDERICK J. HAWRYSH
NATIONAL DEFENCE HEADQUARTERS
ATTN: DMOS 3
101 COL BY DRIVE, OTTAWA
ONTARIO, CANADA K1A 0K2
CANADA

962

1059

*DR. ALLYN HERTZBACH
US ARMY RESEARCH INSTITUTE
ATTN: USARI-PERI-RP
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*REBECCA D. HETTER
NAVY PERSONNEL R&D CENTER
CODE 622
SAN DIEGO, CA 92152-6800

*DR. EDWARD N. HOBSON
11713 LARIAT LANE
OAKTON, VA 22124

*PAUL W. HOLLAND
21-T
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON, NJ 08541

*DR. FRED D. HOLT
COMMANDANT
U.S. ARMY CHEMICAL SCHOOL
ATTN: ATZN-CM-ES (DR. FRED HOLT)
FT MCCLELLAN, AL 36205-5000

*ROBERT F. HOLZ
US ARMY RESEARCH INSTITUTE
PERI-PO
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*MR. CHARLES R. HOSHAW
5920 BROOKVIEW DRIVE
ALEXANDRIA, VA 22310

*LEAETTA M. HOUGH
PERSONNEL DECISIONS RESEARCH
INSTITUTE
43 MAIN ST. SE #405
MINNEAPOLIS, MN 55414

*DR. GLORIA B. HOUSTON
NAVAL EDUCATION TRAINING
& SUPPORT CENTER, PACIFIC (N-7)
SAN DIEGO CA 92132

*WILLIAM K. HUNTSINGER
NAVY OCCUPATIONAL DEVELOPMENT
AND ANALYSIS CENTER
NAVAL MILITARY PERSONNEL COMMAND DET
BLDG. 150, CODE 30
WASHINGTON, D.C. 20374

"I"

*ROBERT A. IRELAND, JR.
U.S. NAVY RECRUITING DISTRICT
5051 RODEO ROAD
LOS ANGELES, CA 90016-4795

"J"

*IAN L. JACKSON
PUBLIC SERVICE COMMISSION
PSC/STAFF DEVELOPMENT BRANCH
L'ESPLANADE LAURIER, WEST TOWER
300 LAURIER AVENUE WEST
OTTAWA, ONTARIO
CANADA K1A 0M7

*DR. JOHN W. JONES
ST. PAUL INSURANCE CO.
RMSD-THE ST. PAUL COMPANY
385 WASHINGTON STREET
ST. PAUL, MN 55102

*KAREN N. JONES
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

*KEN JONES
ROYAL NAVY, UNITED KINGDOM
RNSETT
HMS NELSON
PORTSMOUTH, HANTS PO1 3HH
ENGLAND

*GLORIA JONES-JAMES
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

"K"

1069

*PROF. MICHAEL KASTNER
UNIVERSITY OF THE ARMED FORCES OF THE FEDERAL REPUBLIC OF GERMANY, FB WOW
WERNER-HEISENBERG-WEG 39
8014 NEUBIBERG, WEST GERMANY

*ROBERT S. KENNEDY
ESSEX CORPORATION
1040 WOODCOCK ROAD
SUITE 227
ORLANDO, FL 32803

*DALE J. KERKMAN
TRAINING DEVELOPMENT UNIT
BLDG. 90, NTC
ATTN: NEW TECHNOLOGY OFFICE
GREAT LAKES, FL 60088-5704

*CDR ROBERT H. KERR
323 COLBY DR.
DARTMOUTH, NOVA SCOTIA
CANADA, 32V2B7

*WILLIAM F. KIECKHAEFER
RGI INCORPORATED
1360 ROSECRANS, SUITE D
SAN DIEGO, CA 92106

*DR. MELVIN J. KIMMEL
U.S. ARMY RESEARCH INSTITUTE
ATTN: PERI-RP
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22310

*DR. GRENVILLE C. KING
DIRECTOR
U.S. ARMY MANAGEMENT ENGINEERING TRAINING ACTIVITY
ATTN: AMXOM-MT (DR. KING)
ROCK ISLAND, IL 61299-7040

*DEIRDRE J. KNAPP
U.S. ARMY RESEARCH INSTITUTE
ATTN: PERI-RP
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*DR. EUGENIA M. KOOS
CODE 772, NAS, NORTH ISLAND
COMASWWINGPAC
SAN DIEGO, CA 92135

963

1061

*BRUCE KRAMER
CTB/MCGRAW-HILL
2500 GARDEN RD.
MONTEREY, CA 93940

*DR. LEONARD P. KROEKER
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*HANS J. KUESSNER
U.S. ARMY RESEARCH INSTITUTE
TRL
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*JAMES H. KVICALA
3748 WINKFIELD PL.
COLUMBUS, GA 31909

"L"

*DR. G. J. LAABS
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*DR. ANITA LANCASTER
1325 CALVIN FOREST DRIVE
VIENNA, VA 22180

*RICHARD S. LANTERMAN
COMMANDANT, U.S. COAST GUARD, G-P-1/2
2100 SECOND STREET SW
WASHINGTON, DC 20593

*MS JANICE LAURENCE
HUMAN RESOURCES RESEARCH ORGANIZATION
1100 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

*MS. LINDA C. LAWING
CHARLOTTE MEPS
P.O. BOX 34129
CHARLOTTE, NC 28234

*DR. RICHARD A. LILIENTHAL
U.S. ARMY CIVILIAN PERSONNEL CENTER
PECC-CMP
200 STOVALL STREET
ALEXANDRIA, VA 22332-0300

*LTC LARRY R. LIPPINCOTT
CHIEF, OTEA FIELD OFFICE
ATTN: CSTE-FOB, BLDG T-48
FORT BLISS, TX 79916-7619

*GENERAL HUCK LONG
EDUCATIONAL TESTING SERVICE
1825 I. STREET, N.W.
WASHINGTON, D.C. 20006

*MICHAEL W. LORENZO
3224 YORKTOWN DRIVE
TALLAHASSEE, FL 32312

*ROBERT M. LUND
HQ, USMEPCOM
2500 GREEN BAY RD.
NORTH CHICAGO, IL 60064

"M"

*DR. MILTON MAIER
CENTER FOR NAVAL ANALYSES
4401 FORD AVENUE
P.O. BOX 16268
ALEXANDRIA, VA 22302-0268

*DR. WILLIAM L. MALOY
CHIEF OF NAVAL EDUC AND TRAINING
CODE 00A
NAVAL AIR STATION
PENSACOLA, FL 32508

*CLESSEN J. MARTIN
U.S. ARMY RESEARCH INSTITUTE
PERI-RS
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*JOHN J. MATHEWS
NTEC, CODE 10
NAVAL TRAINING CENTER
ORLANDO, FL 32813-7100

1062

*JOYCE D. MATTSON
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

*PAUL W. MAYBERRY
CENTER FOR NAVAL ANALYSES
4401 FORD AVENUE
P.O. BOX 16268
ALEXANDRIA, VA 22302-0268

*MR. JERRY MCCABE
TEACHING & LEARNING TECHNOLOGIES
3110 WOODCREEK DRIVE
DOWNERS GROVE, IL 60515

*BARBARA L. MCCOMBS
DENVER RESEARCH INSTITUTE
SSRE DIVISION
UNIVERSITY OF DENVER
DENVER CO 80208

*DR. CLARENCE C. MCCORMICK
HQ USMEPCOM/MEPCT-P
2500 GREEN BAY ROAD
NORTH CHICAGO, IL 60064-3094

*PAMELA B. MCCRARY
FEDERAL LAW ENFORCEMENT TRA CENTER
BLDG. 262
GLYNCO, GA 31524

*BARBARA A. MCDONALD
NAVY PERSONNEL R&D CENTER
CODE 51
SAN DIEGO, CA 92152-6800

*JEFFREY J. MCHENRY
PERSONNEL DECISIONS RESEARCH
INSTITUTE
43 MAIN STREET SOUTHEAST
SUITE 405
MINNEAPOLIS, MN 55414

*KENNETH W. MEACHAM
COMMANDANT
USAOMMCS (ATTN: ATSK-ES) DOES
STANDARDIZATION & ANALYSIS DIVISION
REDSTONE ARSENAL, AL 35897-6700

*DR. BARBARA M. MEANS
HUMAN RESOURCE RESEARCH ORGANIZATION
(HUMRRO)
1100 S. WASHINGTON STREET
ALEXANDRIA, VA 22314

*DOUGLAS H. MEBANE
COMMANDANT
DEFENSE LANGUAGE INSTITUTE
DLIELC/LEACT (MR. MEBANE)
LACKLAND AFB, TX 78236-5000

*LARRY L. MELIZA
ARI FIELD UNIT
PO BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*DR. ALBERT MELTER
PERSONALSTAMMAMT DER BUNDESLSEHR
ABT.II/3(OP2), MUDRA-KASERNE
KOLNER STR 262
D-5000 KOLN 90
FEDERAL REPUBLIC OF GERMANY

*CHARLES MICHALK
CNATRA CODE 3122
NAVAL AIR STATION
CORPUS CHRISTI, TX 78419

*DR. ANGELO MIRABELLA
ARMY RESEARCH INSTITUTE
12508 KUHL ROAD
SILVER SPRING, MD 20902

*DR. JAMES L. MITCHELL
MCDONNELL DOUGLAS ASTRONAUTICS CO.
DEPT E422
926 TOEPPERWEIN ROAD
CONVERSE, TX 78109

*ANTHONY E. MIZEN
CINCNAVHOME
HM NAVAL BASE
PORTSMOUTH, HANTS, PO1 3LR
ENGLAND

*DEBORAH MOHR
NAVY PERSONNEL R&D CENTER
CODE 72
SAN DIEGO, CA 92152-6800

*WILLIAM E. MONTAGUE
NAVY PERSONNEL R&D CENTER
CODE 05
SAN DIEGO, CA 92152-6800

*JUDY M. MORACCO
3510 CARLOTTA STREET
PENSACOLA, FLORIDA 32503

*KATHLEEN MORENO
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

*THOMAS MUIR
CHIEF OF NAVAL OPERATIONS
(OP-14)
NAVY DEPT.
WASHINGTON, DC 20350-2000

*DIXIE L. MURDOCK
306 AUSTIN LOOP
FT. BENNING, GA 31905

*KEVIN R. MURPHY
DEPT. OF PSYCHOLOGY
COLORADO STATE UNIVERSITY
FORT COLLINS, CO 80523

"N"

*DELBERT M. NEBEKER
NAVY PERSONNEL R&D CENTER
CODE P72
SAN DIEGO, CA 92152-6800

*WOLFGANG NEEF
UNIVERSITY OF THE ARMED FORCES OF THE
FEDERAL REPUBLIC OF GERMANY, FB WOW
WERNER-HEISENBERG-WEG 39
D-8014 NEUBIBERG, WEST GERMANY

*DR PETER F. NEWTON
NATIONAL SECURITY AGENCY
BOX 1, FANX III
FORT GEORGE G. MEADE, MD 20755-6000

*JUDITH NICHOLS
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

"O"

*DARLENE M. OLSON
19654 FRAMINGHAM DRIVE
GAITHERSBURG, MD 20879
ALEXANDRIA, VA 22333

*CDR. J. E. OLSON
NETPDC, CODE 03
PENSACOLA, FL 32509

*DR. ROLF OTTE
PERSONALSTAMMAMT DER BUNDESWEHR
MUDRAKASERNE
KOLNER STR. 262
5 KOLN 90
GERMANY

*DAVID J. OWEN
DIRECTORATE OF MILITARY
OCCUPATIONAL STRUCTURES
NATIONAL DEFENCE HEADQUARTERS
101 COLONEL BY DRIVE
OTTAWA, ONTARIO
CANADA K1A0K2

"P"

*CAPT MARK L. PARIS, PH.D.
CHIEF, PSYCHOLOGY DIVISION
DIRECTORATE OF MENTAL HEALTH
U.S. DISCIPLINARY BARRACKS
FT. LEAVENWORTH, KS 66027-7100

*DR. JOHN J. PASS
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO CA 92152-6800

*DR. DAVID L. PAYNE
IDAC
3280 PROGRESS DR.
ORLANDO, FL 32826

*DR. EARL PENCE
U.S. ARMY RESEARCH INSTITUTE
ATTN: PERI-RL (DR. EARL PENCE)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*DR. NANCY N. PERRY
CNET CODE 00A2
NAS PENSACOLA
PENSACOLA FL 32514

*DR. NORMAN G. PETERSON
PERSONNEL DECISIONS RESEARCH INSTITUTE
43 MAIN ST. S.E. #405
MINNEAPOLIS, MN 55414

*DR. ALLAN L. PETTIE
USAISCFE
ATIC-ITP
ATTN: DR. ALLAN PETTIE
FORT EUSTIS, VA 23604-5206

*WILLIAM J. PHALEN
AFHRL/MOMM
BROOKS AFB, TX 78235-5601

*EARL R. PHILLIPS
M C AIR STATION CHERRY POINT
EDUCATION OFFICER (CODE TE)
MCAS CHERRY POINT, NC 28533

*VIRGINIA M. PHILLIPS
U. S. ARMY RECRUITING COMMAND
FEDERAL OFFICE BLDG 4-1-150
24000 AVILA ROAD
LAGUNA NIGUEL, CA 92677-1001

*JOHN D. PICKETT, JR.
AFROTC/XPX
MAXWELL AFB, AL 36112-6663

*DON PITTMAN
DOD TEST SPECIALIST
SPOKANE MEPS
U.S. COURT HOUSE
SPOKANE, WA 99201

*CDR EARL H. POTTER
DEPT OF ECONOMICS AND MANAGEMENT
US COAST GUARD ACADEMY
NEW LONDON, CT 06320

*STEPHEN PRESTWOOD
ASSESSMENT SYSTEMS CORPORATION
2233 UNIVERSITY AVENUE, SUITE 310
ST PAUL, MN 55114

*TERRY J. PROCIUK
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
SUITE 600, 4900 YONGE STREET
WILLOWDALE, ONTARIO
CANADA M2N 6B7

*TONY P. PRUCNAL
ALTERNATE TEST CONTROL OFFICER
ARMY EDUCATION CENTER
BUILDING 219
FT MYER, VA 22211-5050

*DR. ALBERT P. PRUETT
NAMTRAGRU NAS MEMPHIS
P.O. BOX 28731
MEMPHIS, TN 38128

"Q"

*RONALD J. QUAYLE
NATIONAL COMPUTER SYSTEMS
1101 30TH ST. N.W., SUITE 500
WASHINGTON, DC 20007

*MAJ JOHN C. QUEBE
9851 FORTUNE RIDGE
CONVERSE, TX 78109

"R"

*EUGENE M. RAMRAS
NPRDC, CODE 01A
SAN DIEGO, 92152-6800

*MARTIN L. RAUCH
MINISTRY OF DEFENSE - P II 4
POSTFACH 1328
53 BONN 1
FEDERAL REPUBLIC OF GERMANY

*LTCOL BRUCE J. REED
COMMANDANT OF THE MARINE CORPS
(CODE TAP-30)
HQ, U.S. MARINE CORPS
WASHINGTON, D.C. 20380-0001

*DR. JAMES A. RIEDEL
NAVY PERSONNEL R&D CENTER
CODE 72
SAN DIEGO, CA 92152-6800

*BARRY J. RIEGELHAUPT
HUMRRO
1100 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

*COL. PAUL T. RINGENBACH
COMMANDER
USAF OCCUPATIONAL MEASUREMENT CENTER/CC
RANDOLPH AFB, TX 78150

*MR. FRANK L. RIPKIN
NAVAL CIVILIAN PERSONNEL COMMAND
NCPC-132 (OP-140F2)
ARLINGTON ANNEX, ROOM G824
WASHINGTON, DC 20350-2000

*DR. GERD W. RODEL
FREIW.-ANNAHMEZENTRALE
DER MARINE
EBKERIEGE-KASERNE
D - 2940 WILHELMSHAVEN
FEDERAL REPUBLIC OF GERMANY

*BOB B. RODGERS
8018 SUNSET PATH COURT
SPRINGFIELD, VA 22153

*KENDALL L. ROOSE
TRAINING AIRWING FIVE
ACADEMIC TRAINING DEPARTMENT
MILTON, FL 32570

*GERARDUS J. ROOZENDAAL
ROYAL NETHERLANDS EMBASSY
OFFICE OF THE ARMY ATTACHE
4200 LINNEAN AVENUE N.W.
WASHINGTON, DC 20008

*HARVEY ROSENBAUM
AMERICAN INSTITUTE FOR RESEARCH
AIR
1055 THOMAS JEFFERSON ST, NW
WASHINGTON, DC 20007

*RODNEY L. ROSSE
PERSONNEL DECISIONS RESEARCH
INSTITUTE
43 MAIN STREET SE #405
MINNEAPOLIS, MN 55414

*PAUL G. ROSSMEISSL
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333-5600

*GREGORY B. RUNYAN
501 TAMPICO BLVD
PENSACOLA, FL 32506

"S"

*DENNIS P. SACCUZZO
13253 DAX COURT
SAN DIEGO, CA 92129

*WILLIAM A. SANDS
NAVY PERSONNEL RESEARCH AND
DEVELOPMENT CENTER (CODE 63)
SAN DIEGO, CA 92152-6800

*GARY G. SARLI
US ARMY RESEARCH INSTITUTE
PO BOX 6057
FORT BLISS, TX 79906-0057

*CLARE N. SCHOFER
14 TRAINING GROUP HQ
CANADIAN FORCES BASE WINNIPEG
WESTWIN, MAN
R2R OTO CANADA

*DR. MARY A. SCHRATZ
NAVY PERSONNEL R&D CENTER
CODE 632
SAN DIEGO, CA 92152-6800

*MARK SCHWARTZ
COMMANDER
U.S. ARMY INTELLIGENCE SCHOOL
ATTN: ATSI-EES (DR. SCHWARTZ)
FT. DEVENS, MA 01433-6301

*DANIEL O. SEGALL
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

*WILLIAM L. SHACKELFORD
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

*NORMAN P. SHERWOOD
ARMY RECRUITING COMMAND
HQ USAREC
ATTN: RO-RS
FT. SHERIDAN, IL 60037

*DR. JOYCE SHIELDS
THE HAY ASSOCIATES
1110 VERMONT AVE., SUITE 710
WASHINGTON, D.C. 20005

*THEODORE M. SHLECHTER
U.S. ARMY RESEARCH INSTITUTE
FT. KNOX FIELD UNIT
STECLE HALL
FT. KNOX, KY 40121-5620

*WAYNE SHORE
PERFORMANCE METRICS, INC.
5825 CALLAGHAN, SUITE 225
SAN ANTONIO, TX 78228

*LT COL LAWRENCE Q. SHORT
AFHRL/MOAO
BROOKS AFB, TX 78235-5601

*DR ALLEN N. SHUB
SHUB & ASSOCIATES
P.O. BOX 572
BUFFALO GROVE, IL 60090-0572

*DR. NORMAN M. SHUMATE
DIRECTORATE OF TRAINING AND DOCTRINE
US ARMY ARMOR SCHOOL
ATTN: ATSB-DOTD
FORT KNOX, KY 40121-5200

*CLARENCE V. SIEGNER
FLEASWTRACENPAC
HARBOR DRIVE & NIMITZ BLVD.
ATTN: DR. C. V. SIEGNER N8B
SAN DIEGO, CA 92147

*LOYD D. SINGLETARY
CHIEF OF NAVAL EDUCATION AND TRAINING
CODE N-5321 NAS
PENSACOLA, FL 32508

*ELIZABETH P. SMITH
COMMANDER
U.S. ARMY RESEARCH INSTITUTE
ATTN: PERI-RS (DR. ELIZABETH SMITH)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*DR. MARGARET J. SMITH
CODE 301, BLDG 2435, RM 137
NETPDC
PENSACOLA, FL 32509

*DUANE L. SOWERS
MEPS, PHOENIX
ATTN: DUANE SOWERS
215 N. 7TH STREET
PHOENIX, AZ 85034

*GARNETT L. SPEARMAN
2328 TRAVIS PINES ROAD
AUGUSTA, GA 30906

*DR GLEN L. SPIVEY
ACSC/CAE
BLDG 1402
MAXWELL AFB, AL 36112-5542

*DR. WILLIAM D. SPRENGER
HQS, TRADOC
ATRO-TA (SPRENGER)
FT. MONROE, VA 23651

*MICHAEL R. STALEY
TEXAS MAXIMA CORP.
8301 BROADWAY, SUITE 212
SAN ANTONIO, TX 78209

*PAUL P. STANLEY II
USAFOMC/OMD (PAUL STANLEY)
RANDOLPH AFB, TX 78150

*BRIAN M. STECHER
(EDUCATIONAL TESTING SERVICE)
815 COLORADO BLVD
SUITE #606
LOS ANGELES, CA 90041

*ROBERT P. STEEL
AFIT/LSB
WRIGHT-PATTERSON AFB, OH 45433

*STAN STEPHENSON
SOUTHWEST TEXAS STATE UNIVERSITY
SAN MARCOS, TX 78666

*CHARLES R. STEWART III
COMMANDER
US ARMY INTELLIGENCE CTR & SCHL
ATTN: ATSI-ES
FORT HUACHUCA, AZ 85613-7000

*MAJ WILLIAM J. STRICKLAND
15880 CLIFFBROOK CT.
DUMFRIES, VA 22026

*LTCOL WILLIAM J. SUBLETTE
COMMANDANT OF THE MARINE CORPS
(CODE TDA-20)
HEADQUARTERS USMC
WASHINGTON, DC 20380-0001

*JAMES B. SYMPSON
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

"T"

*JOESPH S. TARTELL
USAFOMC/OMY
RANDOLPH AFB, TX 78150-5000

*MAURICE M. TATSUOKA
CERL, UNIV OF ILLINOIS
AT URBANA-CHAMPAIGN
220 EDUCATION BUILDING
1310 SOUTH SIXTH STREET
CHAMPAIGN, IL 61820

*CARL J. TAYLOR
AFHRL/MODP
BROOKS AFB, TX 78235-5000

*DAVID M. THISSEN
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF KANSAS
LAWRENCE, KA 66045

*DR. JOHN P. THOMAS
RD#2, BOX 319
DELANSON, NY 12053

*JERRY L. THOMPSON
EDUCATIONAL TESTING SERVICE
PRINCETON, NJ 08541

*DR. RONALD B. TIGGLE
NAVY PERSONNEL R&D CENTER
CODE 631
SAN DIEGO, CA 92152-6800

*SHARON TKACZ
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE.
ALEXANDRIA, VA 22333

*JODY L. TOQUAM
PERSONNEL DECISIONS
RESEARCH INSTITUTE
43 MAIN STREET S.E., SUITE #405
MINNEAPOLIS, MN 55414

*DR. MARVIN H. TRATTNER
US OFFICE OF PERSONNEL MANAGEMENT
1900 E STREET NW
WASHINGTON, DC 20413

*JAMES M. TRIPP
CDR, USA TRAINING SUPPORT CENTER
ATTN: ATIC-ITT (MR. TRIPP)
FORT EUSTIS, VA 23604

*DR. J. TWEEDDALE
TECHNICAL DIRECTOR
NAVY PERSONNEL R&D CENTER
CODE 01
SAN DIEGO, CA 92152-6800
.kp off

"U"

*DR. GEORGE M. USOVA
110 NELSON DRIVE
NEWPORT NEWS, VA 23601

"V"

*DAVID VALE
ASSESSMENT SYSTEMS CORPORATION
2233 UNIVERSITY AVE., SUITE 310
ST PAUL MN 55114

*NICHOLAS H. VANMATRE
NAVY PERSONNEL R&D CENTER
CODE 52
SAN DIEGO CA 92152-6800

*PAUL P. VAN RIJN
ARI (PERI-RL)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*DR. DAVID S. VAUGHAN
DEPT E422, BLDG. 277C/1/F1
MCDONNELL DOUGLAS ASTRONAUTICS CO.
951 HORNET DRIVE
HAZELWOOD, MO 63042

*SALLY I. VICKERS
HQ U.S. WESTERN MEPCOM
BLDG 1808
PRESIDIO OF S.F., CA 94129

*DR. ROBERT VINEBERG
HUMRRO
27857 BERWICK DRIVE
CARMEL, CA 93923

*SUSAN J. VOIDAHL
MILWAUKEE MEPS
1711 BIRCH RD.
KENOSHA, WI 53140

"W"

*DR. LLOYD W. WADE
MARINE CORPS INSTITUTE
SPECIAL PROGRAMS DEPARTMENT
PO BOX 1775
ARLINGTON, VA 22222-0001

*KATHLEEN E. WAGNER
ARMY CONTINUING EDUCATION CENTER
BLDG 2248
AFZG-PTS-TRE
FT. SAM HOUSTON, TX 78234-5000

*DR. MICHAEL P. WAGNER
DYNAMICS RESEARCH CORPORATION
60 CONCORD STREET
WILMINGTON, MA 01887

*DR. HOWARD WAINER
RESEARCH AND STATISTICS GROUP
EDUCATIONAL TESTING SERVICE (21T)
PRINCETON, NJ 08541

*DR. RAYMOND O. WALDKOETTER
US ARMY SOLDIER SUPPORT CENTER
ATTN: ATSG-DSS (DR. R.O. WALDKOETTER)
BUILDING 1
FORT BENJAMIN HARRISON, IN 46216-5060

*CLINTON B. WALKER
ARMY RESEARCH INSTITUTE (PERI-RS)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*GARY H. WANAMAKER
ALLEN CORPORATION
4053 8th Ave. #27
SAN DIEGO, CA 92103

*DR. THOMAS A. WARM
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

*JONATHAN WARREN
2360 EUNICE ST.
BERKLEY, CA 94708

*DR. BRIAN K. WATERS
MANPOWER ANALYSIS PROGRAM
HUMRRO
1100 S. WASHINGTON ST.
ALEXANDRIA, VA 22314

*HARVEY WEBB
HQ U.S. MEPCOM
2500 GREEN BAY ROAD
NORTH CHICAGO, IL 60064-3094

*JOSEPH L. WEEKS
AFHRL/MODJ
BROOKS AFB
SAN ANTONIO, TX 78235

*CAPT TONI G. WEGNER
AFHRL/MOAE
BROOKS AFB, TX 78235-5601

*JOHNNY J. WEISSMULLER
TEXAS MAXIMA CORP
8301 BROADWAY STE 212
SAN ANTONIO, TX 78209

*ROBERT WELLS
NAVY PERSONNEL R&D CENTER
CODE 06
SAN DIEGO, CA 92152-6800

*MAJ JOHN R. WELSH
AFHRL/MOAE
BROOKS AFB, TX 78235-5601

*MAJ KAROL W. J. WENEK
CANADIAN FORCES PERSONNEL APPLIED
RESEARCH UNIT
4900 YONGE STREET, SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

*CHARLES M. WEST
CNET, CODE N-1A
NAS
PENSACOLA, FL 32508

*DOUGLAS C. WETZEL
NAVY PERSONNEL R&D CENTER
CODE 51
SAN DIEGO CA 92152-6800

*MAJ. COLIN P. WHEELER
ARMY SCHOOL OF TRAINING SUPPORT
RAEC CENTRE
BEACONSFIELD, BUCKS
ENGLAND

*LEONARD A. WHITE
U.S. ARMY RESEARCH INSTITUTE
MPRL-PERI-RS
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*MR. PATRICK J. WHITMARSH
U.S. ARMY RESEARCH INSTITUTE FIELD UNIT
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*DR. GERRY L. WILCOVE
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*WOLFGANG WILDGRUBE
STREITKRAEFTEAMT
POSTFACH 20 50 03
D-5300 BONN 2
WEST GERMANY

*RICHARD C. WILLING
US COAST GUARD INSTITUTE
PO SUBSTATION 18 (MVP)
OKLAHOMA CITY, OK 73169-6999

*MAJ F. P. WILSON
CANADIAN FORCES:: DPSRSC
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, CANADA K1A 0K2

*ROBERT J. WILSON
NAVAL MILITARY PERSONNEL COMMAND DET
NAVY OCCUPATIONAL DEVELOPMENT AND
ANALYSIS CENTER
BLDG 150, WASHINGTON NAVY YARD
ANACOSTIA
WASHINGTON, D.C. 20374-1501

*SUSAN T. WILSON
EDUCATIONAL TESTING SERVICE 12-R
PRINCETON, NJ 08541

*DR. WINFORD D. WIMMER
636 BRENTWOOD DRIVE
ANNISTON, AL 36206

*S. WINDLE
ALLEN CORPORATION
10080 CARROLL CANYON RD.
SAN DIEGO, CA 92131

*DR. HILDA WING
409 THAYER PLACE
SILVER SPRING, MD 20910

*DR. MARTIN F. WISKOFF
NAVY PERSONNEL RESEARCH
& DEVELOPMENT CENTER
CODE 06
SAN DIEGO CA 92152-6800

*DR. JOHN WOLFE
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

*MR. ROBERT J. WOLFINGTON
COMMANDER
U.S. ARMY AVIATION CENTER
ATTN: ATZQ-TD-IV-E (MR. WOLFINGTON)
FORT RUCKER, AL 36360-5000

*JO ANNA WOOD
DEPARTMENT OF PSYCHOLOGY
SOUTHERN ILLINOIS UNIVERSITY
CARBONDALE, IL 62901

*DARRELL A. WORSTINE
SOLDIER SUPPORT CENTER-NCR
200 STOVALL STREET
ALEXANDRIA, VA 22332-0400

*GREGG J. WRIGHT
BOOZ, ALLEN & HAMILTON
7315 WISCONSIN AVE - 1100W
BETHESDA, MD 20814

"Z"

*TIMOTHY C. ZELLO
COMMANDER
U.S. ARMY ORDNANCE CENTER & SCHOOL
ATTN: ATSL-DES-E (TIMOTHY ZELLO)
ABERDEEN PROVING GROUND, MD 21005-5201

*COLONEL GERRY A. ZYPCHEN
NATIONAL DEFENCE HEADQUARTERS
ATTN: DMOS
2781 SPRINGLAND DRIVE
OTTAWA, ONTARIO
CANADA, K1V 9X2

696

AUTHOR INDEX

AUTHOR	PAGE
ABRAHAMS, N.M.	503
ADAMS, J.	210
AMYOT, CAPT K.A.	462
ANDERSON, J.W.	153
ANGUS, CAPT R.J.	547
ANSBRO, T.M.	701
ARABIAN, J.M.	300,603
ARCHER, BG C.J.	xxvi
ARMIJO, L.	559
ATWATER, L.	113
ATWOOD, N.K.	798
BAKER, H.G.	327,567,780
BALLENTINE, R.D.	306
BANKS, J.H.	916
BARGE, B.N.	582,891
BART, W.M.	707
BEARDEN, R.M.	317
BELL, 1LT J.M.	228
BILLS, MAJ C.G.	666
BLACK, B.A.	786
BLACKHURST, CPT J.L.	306,327
BLYTH, D.M.	741
BOLDOVICI, J.A.	444,826
BONCZAR, T.P.	491
BONDARUK, J.	543
BORMAN, W.C.	367
BRANDT, D.A.	730
BRIDGEMAN, B.	337
BROWN, G.E.	650
BURT, J.A.	352
BUSH, B.J.	838
BUTLER, W.G.	204
CANTOR, K.A.	832
CECIL, S.	486
CELESTE, J.F.	526
CHANG, F.R.	553
CHATFIELD, R.E.	222
COOPER, M.	621
CORPE, V.A.	885
CORY, C.H.	775
COSTELLIC, MAJ M.	194
CRAWFORD, A.	113
CRAWFORD, K.S.	73
CROWLEY, N.R.	159
DANSBY, MAJ M.R.	531
DIAMOND, E.E.	570
DILLA, CPT B.L.	118,216
DOCKSTADER, S.L.	61
DOHERTY, L.M.	480
DOHERTY, W.J.	928,931
DRISKILL, W.E.	198
DUNCAN, CPT R.E.	713
DUNLAP, W.P.	107
DUNNETTE, M.D.	885
EATON, N.K.	769
EBENRETT, H.	251
EDDINS, J.M.	13
EKSTROM, A.G.	792
EKWALL, R.W.	943
ELLINGSWORTH, M.E.	420
ELLIS, MAJ R.T.	547
FAIRBANK, B.A.	136
FINE, B.J.	843
FISHER, G.P.	897,900
FORSYTHE, T.K.	928
FRENCH, C.M.	809
FREY, R.L.	736
GARCIA, S.K.	262
GAST, I.F.	147
GIRDLER, K.	130
GOEHRING, D.J.	239,655
GOLDMAN, L.A.	358,537
GORMAN, LTC C.D.	198
GOTTESMAN, A.M.	165
GRAHAM, S.E.	826
GRIEGER, L.W.	644
HANLON, J.P.	759
HANSER, L.M.	300
HARDING, F.D.	90,467
HARDY, G.R.	747
HARDY, LT D.L.	198
HARMAN, J.	404
HART, R.J.	655
HEISEY, J.G.	855
HERTZBACH, A.	587
HETTER, R.D.	503
HIGHTOWER, CPT J.M.	375
HOLLAND, P.W.	282
HOUGH, L.M.	582,891,900
HOUSTON, J.S.	885,891
IBSEN, CPT K.A.	531
JOHNSON, D.M.	615
JOHNSON, J.H.	398
JOHNSON, R.M.	102,587
JONES, J.W.	637
JONES, K.N.	352
JONES, LCDR K.	341
JOYNER, J.N.	323
KAMP, J.D.	891
KANTOR, J.E.	680

AUTHOR INDEX

AUTHOR	PAGE
KASTNER, M.	438
KENNEDY, R.S.	107,398
KERKMAN, D.	393
KERR, CDR R.H.	88
KERSHAW, S.W.	288
KIECKHAEFER, W.F.	29,38,55
KIMMEL, M.J.	188
KING, G.C.	906
KNAPP, D.J.	102,587
KOBRICK, J.L.	843
KOFFMAN, N.	559
KOROTKIN, A.L.	621
KROEKER, L.PL	317
LAABS, G.J.	317,780
LANCASTER, A.R.	849
LANE, N.E.	107
LARSON, G.E.	233
LAU, A.	769
LAURENCE, J.H.	861
LEWANDOWSKI, CPT F.	216
LILIENTHAL, R.	897,900
LOCKHART, J.M.	615
LOWE, CPT J.K.	375
MACPHERSON, D.	515
MAIER, M.H.	311
MALOY, W.L.	124
MATHEWS, J.J.	809
MATTSON, J.D.	503
MEANS, B.	855
MELIZA, L.L.	661
MERKLE, P.J.	398
MILLER, C.R.	599
MIRABELLA, A.	520
MITCHELL, J.L.	265,271,276
MIZEN, LCDR A.E.	341
MOHR, D.A.	67
MORABITO, CPT M.A.	118
MORENO, K.E.	29,38,55
MORRISON, R.F.	222
MUELLER, H.A.	686
MUMFORD, M.D.	90,467
McBRIDE, J.R.	43
McCOMBS, B.L.	331
McCORMICK, C.	627
McDONALD, B.	450
McGUE, M.K.	891
McHENRY, J.J.	879
McLAIN, LTC F.	130
McLAUGHLIN, D.H.	730
NEBEKER, D.M.	78,82
NEEF, W.	438
NEWTON, P.	543
NICHOLS, J.J.	931
NIGAM, A.	855
NULLMEYER, R.T.	666
OLIVIER, L.F.	420
OLSON, D.M.	367
PARIS, M.L.	650
PASS, J.J.	565
PENCE, E.C.	922
PERRIN, B.M.	265,271
PERRY, N.N.	124
PETERSON, N.G.	867,873
PETTIE, A.L.	821
PFEIFFER, G.J.	420
PHALEN, W.J.	276,414
PIERCE, J.E.	537
PITZ, G.F.	171
PLISKE, R.M.	102
POST, T.	337
POTTER, E.H.	473
QUEBE, J.C.	695
RAFACZ, B.A.	23,567
REE, M.J.	672
RICHARDS, J.D.	210
RIEDEL, J.A.	78
RIEGELHAUPT, B.J.	491
RIMLAND, B.	233
RIPKIN, F.L.	486
ROBERSON, K.	559
ROSENBAUM, H.	815
ROSSE, R.L.	873
ROSSMEISSL, P.G.	609,730
RUMSEY, M.G.	147
SACCUZZO, D.P.	233
SACHS, S.A.	632
SAKO, S.	763
SANDS, W.A.	19,567
SARLI, G.G.	615
SCHNEIDER, E.F.	861
SCHRATZ, M.K.	34
SCHWARTZ, M.M.	725
SEGALL, D.O.	29,38,55
SELLMAN, W.S.	852
SHACKELFORD, W.L.	937
SHANE, G.S.	426
SHIFLETT, S.	621
SHLECHTER, T.M.	444
SHORT, LTC L.O.	375
SHUB, A.N.	432

AUTHOR INDEX

AUTHOR	PAGE
SLAUGHTER, LTC W.J.	763
SMITH, E.P.	593
SMITH, M.G.	398
SPRENGER, W.D.	753
STALEY, M.R.	414
STEEL, R.P.	426
STEPHENSON, S.D.	346,456
STERN, B.M.	632
STICHT, T.	559
TARTELL, J.S.	228
TATSUOKA, M.	1
TAYLOR, C.J.	306
THISSEN, D.	294
TIGGLE, R.B.	23
TKACZ, S.	177
TOQUAM, J.L.	879,885
TRATTNER, M.H.	509
TYERMAN, D.	88
USOVA, G.M.	804
VALE, C.D.	7
VAN RIJN, P.	381
VAUGHAN, D.S.	265,271
VINEBERG, R.	323
WAGNER, M.	194
WAINER, H.	288
WALDKOETTER, R.O.	387
WALKER, C.B.	497
WALKER, L.	832
WARM, T.A.	142
WARREN, J.	543
WATERS, B.K.	719
WEEKS, J.L.	90,467
WEGNER, T.G.	672,676
WEISSMULLER, J.J.	414
WELSH, J.R.	676
WENEK, MAJ K.W.J.	245
WETZEL, C.D.	43
WHEELER, MAJ C.P.	408
WHITE, L.A.	147,632
WHITEHILL, B.	450
WHITMARSH, P.J.	912
WILCOVE, G.L.	183
WILDGRUBE, W.	96
WILKES, R.L.	107
WILSON, MAJ F.P.	256
WING, H.	582,632,769
WISE, L.L.	300,730
WOLFE, J.H.	49
WOOD, J.A.	171
WUEBKER, L.J.	637
YADRICK, R.M.	265,271
YATES, L.G.	515